网络空间信息传播建模分析

蔡皖东　编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

P2P 网络、社交网络、微博网络、网络论坛等网络信息交流平台极大方便了人们的信息共享和交流，同时也带来网络信息安全等方面的挑战，促使人们对网络信息传播机理的研究，探寻其中的信息传播特性和内在规律，为应对网络信息安全挑战提供科学依据和解决方案。

本书主要采用数学建模方法对 P2P 网络、社交网络、微博网络、网络论坛四种网络信息交流平台的信息传播特性和规律进行建模分析和研究，其研究成果可以为优化网络平台结构、改善网站服务功能、正确引导网络舆论、抑制不良信息传播等提供技术方案和参考。全书分为 6 章，分别介绍了网络建模基本理论、P2P 网络特定信息传播模型、社交网络用户关系模型、微博网络用户转发模型以及网络论坛信息传播模型等内容。

本书可作为从事相关研究工作的科技人员参考书以及研究生教材。

图书在版编目（CIP）数据

网络空间信息传播建模分析 / 蔡皖东编著. —北京：电子工业出版社，2017.3

ISBN 978-7-121-30947-2

Ⅰ. ①网… Ⅱ. ①蔡… Ⅲ. ①网络传播－研究 Ⅳ. ①G206.2

中国版本图书馆 CIP 数据核字（2017）第 029838 号

策划编辑：窦 昊
责任编辑：窦 昊
印 刷：
装 订：
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编：100036
开 本：787×980 1/16 印张：15.5 字数：357 千字
版 次：2017 年 3 月第 1 版
印 次：2017 年 3 月第 1 次印刷
定 价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010)88254466，douhao@phei.com.cn。

前　言

互联网被认为是继报纸、杂志、广播、电视四大传统媒体之后的第五媒体，随着互联网的发展，Web 网站、网络论坛、电子邮件等传统的信息交流方式已经难以满足广大网民不断增长的信息交流需求，不断出现了一些新型的网络信息交流方式和平台，如 P2P（Peer to Peer）网络、社交网络、微博、微信等，极大方便了人们的信息共享和交流，用户规模都是数以亿计。任何网络信息交流平台都具有双重性，在方便人们信息交流的同时，也带来网络信息安全等方面的挑战。

P2P 网络是一种基于对等计算模式的分布式网络系统，在互联网应用中占有重要的地位，尤其是 P2P 文件共享系统应用最为广泛，它们所产生的网络流量占有互联网流量的很大比例。P2P 文件共享系统在方便人们共享信息的同时，也带来了网络流量管理、知识产权保护、网络信息安全等方面的问题，特别是网络信息安全问题比较突出，蠕虫、病毒、木马等恶意代码借助 P2P 网络的强大传播能力，可以在一夜之间感染成千上万台机器，严重地影响着网络安全；在 P2P 网络上充斥着大量的色情、暴力、迷信、反华宣传等不良信息，给社会和谐稳定、网络文化安全乃至国家安全带来极大的危害。

在线社交网络是一种新型的互联网应用，将社会学中的社交网络概念应用于互联网中，成为人们网上交往和信息交流的热门工具和平台，受到广大网民的欢迎。由于在线社交网络是围绕着用户来建立和组织的，因此可以利用在线社交网络建立的朋友关系进行广告宣传、产品推介、信息交流、活动联络等，在促进人们的社会交往和信息交流方面起到积极的作用，同时也会被不法的组织和人员利用进行非法联络和谣言传播，例如，在国内外发生的暴力恐怖事件中，恐怖组织利用在线社交网络进行谣言散布和非法联络。

微博是一种集成化、开放式的互联网社交服务平台，用户通过微博平台以 140 字左右的文字发布信息，实现即时分享。用户可以根据自己的兴趣爱好，选择关注其他用户，构建自己的关注网络。微博网络作为一种特殊的社交网络，用户不但可以有选择地连接感兴趣的用户，关注其信息，而且也可以被其他用户相互连接，交流信息，具有社交网络和媒体网络的双重特性。微博作为新兴的社交媒体，越来越受重视。例如，美国总统特朗普在竞选期间通过 Twitter 与选民进行交流互动，赢得了更多选民的支持；国内的 CCTV、人

民日报、新华通讯社等主流媒体报刊都在新浪微博平台上开通了官方微博。微博网络作为一种特殊的社交网络，存在着与社交网络相类似的网络安全问题，如非法联络、传播谣言、煽动闹事等，容易引发社会群体事件，给社会和谐稳定以及国家安全带来极大的威胁。

网络论坛属于传统的网络信息交流平台，具有多元化、开放性、匿名性及互动性等特点，成为广大网民发表言论、获取信息的重要网络平台。网民在网络论坛上就某些社会问题或公共事务表达不同看法和观点，成为网络舆论的主要来源地。随着网络舆论对社会和公众影响的不断增大，出现了以网络炒作为营生的网络公关公司、网络推手、网络水军等，以各种手法和名目炮制网络热点事件，捧红各色人物，形成虚假的网络舆情，产生错误的舆论导向，危及政府的公信力，容易引发社会群体性事件。

各种网络信息交流平台所带来的网络信息安全问题，促使人们对网络信息传播机理进行研究，探寻其中的信息传播特性和内在规律，为应对网络信息安全挑战提供科学依据和解决方案。

本书主要采用数学建模方法对 P2P 网络、社交网络、微博网络、网络论坛四种网络信息交流平台的信息传播特性和规律进行建模分析和研究，其研究成果可以为优化网络平台结构、改善网站服务功能、正确引导网络舆论、抑制不良信息传播等提供技术方案和参考。

全书分为 6 章，第 1 章为绪论，主要介绍 P2P 网络、社交网络、微博网络以及网络论坛的基本概念和信息传播模式等；第 2 章为网络建模基本理论，介绍网络的图表示方法、复杂网络基本理论、经典网络模型、传播动力学模型等内容；第 3 章为 P2P 网络特定信息传播模型，介绍 P2P 网络测量模型、P2P 信息传播动力学模型、P2P 特定信息传播特性、P2P 特定信息传播控制等内容；第 4 章为社交网络用户关系模型，介绍社交网络信息传播模型、社交网络关系强度模型、社交网络弱连接分析、社交网络用户关系预测、社交网络意见领袖识别等内容；第 5 章为微博网络用户转发模型，介绍微博用户转发特性、微博转发行为预测、微博转发特性预测、微博转发峰值分析、微博意见领袖识别等内容；第 6 章为网络论坛信息传播模型，介绍网络论坛舆情形成模型、网络论坛意见领袖识别、网络论坛水军热帖检测、网络水军账号检测等内容。

本书是我们团队多年来相关研究工作的总结，丁军平博士、张胜兵博士、罗知林博士、徐会杰博士等参与了相关研究工作，并为本书的撰写做出了贡献。因此，本书是团队集体劳动和智慧的结晶。如果本书能够对从事相关研究工作的科技人员以及研究生起到参考借鉴作用的话，作者的目的便达到了。不足之处敬请广大读者批评指正。

最后，感谢西北工业大学教材专著出版基金对本书的大力资助。

作　者

于西北工业大学

目　录

第 1 章

绪　　论

1.1　引言

互联网被认为是继报纸、杂志、广播、电视等四大传统媒体之后的第五媒体，随着互联网的发展，Web 网站、网络论坛、电子邮件等传统的信息交流方式已经难以满足广大网民不断增长的信息交流需求，不断出现了一些新型的网络信息交流方式和平台，如 P2P（Peer to Peer）网络、社交网络、微博、微信等，极大地方便了人们的信息共享和交流。

任何网络信息交流平台都具有双重性，在方便人们信息交流的同时，也带来网络信息安全等方面的挑战。同时，也促使人们对各种网络信息交流平台的信息传播机理进行深入研究，探寻其中的信息传播特性和内在规律，为应对网络信息安全挑战提供科学依据和解决方案。

在这一领域的研究中，国内外学术界提出了多种方法，从不同的角度对网络信息传播机理进行研究，数学建模是其中一种重要的研究方法，主要运用复杂网络理论、传播动力学以及统计学等方法对特定的网络系统进行抽象描述和分析，试图揭示该系统信息传播特性和规律，通过实际数据分析或系统仿真实验等方法来验证模型和算法的正确性和有效性。

本书主要采用数学建模方法对 P2P 网络、社交网络、微博网络、网络论坛等四种网络信息交流平台的信息传播特性和规律进行建模分析和研究。由于每种平台的系统构成、信息传播模式都有所不同，因此首先对这四种网络信息交流平台作一简单介绍。

1.2　P2P 网络概论

P2P 网络是一种基于对等计算模式的分布式网络系统，改变了传统的客户机/服务器（C/S）模式，具有自组织性、可扩展性、鲁棒性、容错性以及负载均衡等优点，在互联网应用中占有重要的地位，用户数以亿计。其中，P2P 文件共享系统应用最为广泛，它们

所产生的网络流量占有互联网流量的很大比例。根据有关统计数据，在欧洲顶级骨干网流量中，P2P 网络流量占有 60%以上的比例；在南美地区的互联网中，超过 50%的网络流量是由 P2P 文件共享业务产生的；在中国的互联网中，P2P 网络流量所占的比例高达 70%，其中 BitTorrent、迅雷和 eMule 三种 P2P 文件共享软件所产生的流量分别占据前三位。可见，P2P 网络流量已经成为各国和地区互联网流量的重要组成部分。

1．P2P 网络特点

在学术界和工业界，对 P2P 网络没有给出一个统一的严格定义，不同的机构对 P2P 网络给出了不同的定义。总体上看，P2P 网络应具有如下的特点。

（1）非中心化。P2P 网络没有类似 C/S 模式的中心服务器，网络资源和服务分散在所有节点上，信息传输和服务实现都直接在节点之间进行，可以无须中间环节或服务器的介入；

（2）可扩展性。在 P2P 网络中，随着节点数的增加，不仅服务的要求增加了，系统的资源和服务能力也在同步扩充，始终能够满足用户需要；

（3）健壮性。由于 P2P 网络的服务是分散在各个节点之间的，部分节点或网络遭到破坏对其他部分的影响较小，具有耐攻击、高容错的优点；

（4）自组织性与自治性。P2P 网络中的节点可以在没有仲裁者的情况下自己维护网络的连接和性能，其网络拓扑会随着节点的加入、离去或失效而重新组织；

（5）负载均衡。由于 P2P 网络中的资源分布在多个节点，不会出现传统网络中少数节点负载过重而大部分节点资源没有充分利用的情况。

2．P2P 网络类型

根据网络拓扑组织形式，可以将 P2P 网络分为以下四种类型。

（1）集中式 P2P 网络。集中式 P2P 网络以目录服务器为中心形成星型结构。节点维护自身资源，通过在目录服务器上注册完成加入 P2P 网络的过程，并将自身信息和资源信息上传到目录服务器上，目录服务器负责节点查找和资源搜索。节点之间的交互与资源共享等行为是以对等模式直接在节点之间进行的，而无须经过目录服务器。集中式 P2P 网络是 P2P 网络的初始形态，网络拓扑结构简单，易于部署和管理，可以避免资源传递时对服务器所产生的网络流量压力。由于节点查找和资源搜索是通过目录服务器集中完成的，因此存在单点失效和性能瓶颈问题。典型的集中式 P2P 网络主要有 Napster、Aimster、Softwax、iMesh 等。

（2）全分布非结构化 P2P 网络。全分布非结构化 P2P 网络完全按照对等计算模式自组织形成，取消了目录服务器，解决了网络结构中心化问题，扩展性和容错性较好。节点随机接入网络，并与邻居节点通过端到端连接构成逻辑覆盖网络，这种结构能够很好地适

应节点频繁加入、退出及失效的动态环境。资源搜索通过相邻节点广播接力传递，每个节点记录搜索轨迹，防止搜索环路产生。由于资源分散在各个节点，整个网络没有统一的资源管理方式，这就给资源搜索带来一定的困难，控制信息消耗了大量带宽并容易造成网络拥塞，由于没有确定的网络拓扑结构，无法保证资源搜索效率。典型的全分布非结构化P2P网络主要有Gnutella、Freenet等。

（3）全分布结构化P2P网络。全分布结构化P2P网络的拓扑结构是通过分布式哈希表（Distributed Hash Table，DHT）协议进行控制的，资源也由DHT协议精确发布到特定的节点上。这种网络结构的优点是资源定位准确并且能保证一定的效率，具有着良好的可扩展性和搜索性能，适用于对可用性要求高的系统。但结构化P2P网络的维护相对复杂，通常只支持精确匹配，对复杂搜索条件支持较差。典型的全分布结构化P2P网络主要有Chord、Can、Pastry、Tapestry等。

（4）混合式P2P网络。混合式P2P网络吸取了集中式P2P网络和全分布非结构化P2P网络各自的优点。选择性能较高的节点作为超级节点，充当其他普通节点的目录服务器。这些超级节点由P2P网络系统动态选择和组织，不会给P2P网络系统带来单点失效问题。搜索时的路由消息仅在超级节点之间转发，搜索完成后，再将搜索结果返回给相应的叶子节点。混合式P2P网络是一种层次式网络，超级节点之间构成高速转发层，可采用完全对等的方式组织，超级节点和普通节点构成星型网络。这种拓扑将集中式拓扑的易管理性与分布式拓扑的可扩展性有机结合起来，在异构P2P网络环境下是一种较好的选择。典型的混合式P2P网络主要有BitTorrent、KaZaA和eMule等。

3．P2P网络应用

P2P网络以应用为驱动力，在一些领域得到了较好的应用。下面是一些典型的P2P网络应用。

（1）文件共享。在传统方式中，文件提供者将待交换的文件上传到网站服务器，下载者从服务器上下载。这种下载方式在用户多、文件大时，服务器容易过载，下载速度难以得到保证。利用P2P网络技术，计算机之间可以直接交换数据和文件，而不需要借助服务器的中转。Napster是世界首个P2P文件共享系统，为了满足人们对自由共享和交换MP3音乐的需求而开发的，进而引发了P2P网络技术革命，同时也引起了MP3音乐版权纠纷。典型的P2P文件共享系统还有BitTorrent、eMule、KaZaA等。

（2）视频组播。视频组播对带宽要求很高，传统基于C/S模式的视频组播系统由于受到服务器出口带宽的限制，系统可扩展性比较差。在基于P2P的视频组播系统中，只有少数节点从服务器直接获取数据，更多节点在从其他节点处获得数据的同时，也向其他节点提供数据。采用P2P网络技术构建的视频组播系统充分利用了节点之间的可用带宽，

使得系统可扩展性大为提高。典型的视频组播系统主要有 PPLive、PPStream、UUSee 等，它们都拥有大量的用户群。

（3）分布式计算。P2P 网络技术应用于分布式计算时，节点不但接收计算任务，还可以再搜索其他空闲节点并把接收到的任务分发下去。中间结果层层上传，最后到达任务分发节点。这种方式可以合理地整合闲散的计算资源，使总体计算能力得到大规模的提升。斯坦福大学的 Folding@home 项目通过这种分布式计算系统来研究蛋白质折叠、误折、聚合及由此引起的相关疾病。

（4）实时通信。Skype 语音通信软件是一种高质量的 IP 电话系统，它完全采用了 P2P 网络技术，当用户之间需要语音通信时，Skype 在覆盖网中找出一条当前带宽最大的覆盖网通路，通过多跳转发方式进行数据传送。由于 Skype 的出现和高速发展，越来越多的用户转向 IP 电话，对传统的电信业造成了较大的冲击。

（5）数据存储。在基于 P2P 的数据存储系统中，通过 P2P 网络将数据存放在多个 P2P 节点上，而不是专用服务器上。这样不仅可以减轻服务器负担，还可以提高数据存储的可靠性和传输速度。P2P 数据存储系统是以数据的可用性、持久性、安全性为目标，致力于海量数据管理。典型的 P2P 数据存储系统主要有 OceanStore、CFS 和 Granary 等。

4．P2P 文件共享

P2P 文件共享系统是最常用的一种 P2P 网络应用，用户可以利用互联网中特定的 P2P 文件共享系统（如 BitTorrent、迅雷和 eMule 等）来共享各种文件。

在实际应用中，当一个用户提供某个文件共享时，首先需要使用特定 P2P 文件共享软件（简称 P2P 软件）来制作该文件的“元信息”（如 BitTorrent 中的种子文件），在种子文件中包含有提供该文件共享的节点列表查询的可用超级节点（目录服务器）地址和文件标识信息；然后通过网页或 P2P 软件内部发布功能（如 BitTorrent 中的种子市场）等方式发布其种子文件。

当其他用户使用特定 P2P 软件来下载某个特定文件时，首先需要获得该文件的种子文件，P2P 软件将根据种子文件来访问可用的超级节点，找出提供该文件共享的节点列表，利用这些节点完成该文件的快速下载，实现文件共享和信息传播。

5．P2P 网络安全

P2P 文件共享系统在方便人们共享信息的同时，也给网络管理和安全带来了新的问题。在网络流量方面，由于 P2P 网络流量在互联网流量中所占比例比较大，很多电信运营商屡屡指责 P2P 应用抢占网络带宽，甚至联合起来对 P2P 网络流量进行封堵；在知识产权保护方面，由于 P2P 文件共享系统存在知识产权保护的问题，美国 RIAA 等版权组织长期以来针对各种 P2P 文件共享系统所进行的抵制和起诉活动从未中断过；在网络安

全方面，蠕虫、病毒、木马等恶意代码借助 P2P 网络的强大传播能力，可以在一夜之间感染成千上万台机器，严重地影响着网络安全；在 P2P 网络上充斥着大量的色情、暴力、迷信、反华宣传等不良信息，给社会和谐稳定、网络文化安全乃至国家安全带来极大的危害。

因此，研究 P2P 网络信息传播机制是非常必要的，通过对 P2P 信息传播过程的建模分析，研究 P2P 信息传播特性和内在规律，寻找影响 P2P 特定信息传播的关键因素，为抑制 P2P 不良信息传播提供科学依据和解决方案。

1.3 社交网络概论

随着互联网的发展，出现了一种称为在线社交网络的新型互联网应用，将社会学中的社交网络概念应用于互联网中，成为人们网上交往和信息交流的热门工具和平台，受到广大网民的欢迎并得到广泛应用。

中国互联网络信息中心（CNNIC）于 2016 年 1 月发布的《第 37 次中国互联网络发展状况统计报告》显示，截至 2015 年 12 月，中国网民规模达到 6.88 多亿人，互联网普及率达到 50.3%。统计数字表明，我国网民的互联网沟通交流方式不断变化，社交网络、微博等社交应用的网民数达到 5.3 多亿人，使用率为 77%；电子邮件的网民数达到 2.58 多亿人，使用率为 37.6%；网络论坛（BBS）的网民数达到 1.19 多亿人。使用率为 17.3%，以上统计数据表明，网民使用微博、在线社交网络等新型信息交流平台进行信息交流已经成为主流方式。

在线社交网络是社会学的社交网络原理在互联网中的一种创新性应用，为了认识在线社交网络的信息传播机制，首先需要了解社会学的社交网络概念、原理和特性，为研究在线社交网络的基本特性和交流机制提供基础。

1．社交网络基本概念

根据维基百科的定义，社交网络是指个人之间的关系网络，即社交网络是社会个体成员之间因为互动而形成的相对稳定的关系体系，它关注的是人们之间的互动和联系，因为社会互动将影响人们的社会行为。

一个社交网络是由多个社会个体和他们之间的关系组成的集合，可以用网络节点来表示社交网络中的个体，用节点之间的连线表示个体之间的关系，也就是使用几何拓扑学方法来描述社交网络的结构。

社交网络概念强调社会中每一个成员和其他成员之间都存在着或多或少的关系，他们共同构成了这个网络。研究人员通过数学建模的方法来研究社交网络中成员之间的关系，

并对它们之间的联系和特点进行分析，并且着重于个体之间的交互关系，期望能找出社交网络中个体关系的内在联系。

（1）节点。社交网络中的节点表示构成网络的社会个体，而社会个体之间的各种社会关系可以定义为社交网络中的边。在社交网络中，节点也可以称为行动者，节点可以表示任何一个社会个体。例如，节点可以是社交网络中的一个人、小组、单位、组织、企业等，甚至也可以是一个城市、国家等，当然也包括网络中的一个虚拟社区以及组成这个虚拟社区的成员。

（2）关系。每个节点之间的联系依赖于各种关系。这些关系包括人与人之间的情感关系，如喜欢和厌恶；资本之间的交换关系，如商业交易和物资流动；无形资源的转换关系，如朋友交往和信息交换；生物意义上的关系，如配偶和子女等。这些关系都是社交网络的研究对象。

在节点之间还存在着“多元关系”，也就是连接。例如，两个节点之间可能同时存在同事关系、邻里关系等。并且对一个节点来说，有些节点比较重要，而有些节点则无足轻重，因此按联系的紧密程度可分为强连接和弱连接。一个节点与其关系较为亲密的、特别的、经常交互的社会关系之间形成的连接是强连接。与之相反，节点与其不紧密交互或是间接交互的社会关系之间形成的连接是弱连接。

从个人或社会角度来看，弱连接扮演着信息交流、资源传递的重要角色，因此是社交网络研究的一个重点。对于强连接来说，由于彼此之间有着相似的社会背景、知识经验、生活圈子以及相互有很深的了解，因此结构几乎重合，存在着相当多的冗余数据。而弱连接所提供的信息或知识有比较大的差异性，例如那些久不见面的人，他们可能掌握很多自己并不了解的信息。正是由于这些弱纽带的存在，信息才能在不同的圈子中交流。弱连接在社交网络中的虚拟社区之间构建起某种形式的桥梁，可以传递更多的信息和知识。

2．社交网络理论基础

社交网络中有两个重要的理论基础：六度分隔理论和 150 法则。

1）六度分隔理论

该理论是由美国著名社会心理学家 Stanley Milgram 于 20 世纪 60 年代提出的。1967 年，Stanley Milgram 为了描绘一个连结人与社区的人际关系网，设计了一个连锁信实验，将一封含有波士顿一位股票经纪人名字的信件随机发给 160 个人，并要求将此信件转交给比较接近那个股票经纪人的熟人，然后依次办理。最终，这些信件一般经过五六步就能送到该股票经纪人的手中，这就是“六度分隔”现象。简单地说就是，“一个人和任何一个陌生人之间所间隔的人不会超过六个，也就是说，最多通过六个人中的一个人就能够认识任何一个陌生人”。

六度分隔理论说明了社会中人与人之间普遍存在着弱连接，绝对没有联系的社会个体

是不存在的。这种弱连接在社会交往中往往发挥着非常强大的作用，通过弱连接可以间接找到地理位置相距很远的个体，并能够建立联系，从而扩大自己的人际网络，人与人之间的社会距离变得越来越近。因此六度分隔现象又被称为小世界现象。

后来，康奈尔大学的 Jon Kleinberg 用一个二维网格数学模型来描述这个问题，研究证实了小世界现象普遍存在于现实中的各种网络。

2）150 法则

起源于欧洲的“赫特兄弟会”是一个自给自足的农民自发组织，他们有一个不成文的严格规定：每当聚居人数超过 150 人的规模，他们就把这个群体变成两个，再各自发展。因为他们认为，“把人群控制在 150 人以下似乎是管理人群的最佳和最有效的方式”。150 法则成为普遍公认的“可以与之保持社交关系的人数的极限”，也是网络社会化进程中人们应该遵守的规则。

3．社交网络基本特性

（1）社交网络大小。社交网络大小是指组成社交网络的节点数量，又被称为“网络广度”。在社交网络研究中，通常将一个较大的社交网络分解成若干个规模较小的社交网络来处理。

（2）社交网络密度。社交网络密度用于衡量社交网络中各个节点之间联系的紧密程度，用一个社交网络中的实际关系数与最多可能拥有的关系数的比例来度量。

（3）社交网络同质性。社交网络同质性是指构成社交网络的各个节点的背景相似程度。构成同一社交网络的行动者背景越相同，其同质性也越高、越简单，呈单一性；反之，网络的同质性越低、越复杂，呈多元性。同质性越高，网络密度就有可能越大，反之，网络密度就越小。

4．在线社交网络类型

在线社交网络是社交网络在互联网中的实现和应用，人们通过在线社交网络平台，以互联网为媒介进行交友活动和信息交流。在早期的互联网中，在线社交网络就有了相应的雏形。例如，在网络上互发 E-mail 的用户之间就构成了在线社交网络。用户参与到一个网络或虚拟社区中，发布他们的照片或作品，还可以建立与朋友的链接。在线社交网络为维护社会关系和增强信息交流提供了平台，用户可以发现与他们有共同爱好的朋友，还可以下载由其他用户提供的信息。因此，在线社交网络是围绕着用户来建立和组织的。在目前的在线社交网络中，主要有如下几种类型。

（1）在线社区。在线社区通常以个人主页的形式出现，这些主页归属于大型的社区网站。在基于在线社区构建的网络中，节点可以是个人主页，边则是主页间的“友情链接”，整个网络可以抽象成一个有向无权图。在线社区网中的节点也可以是用户，而边则

是用户间的好友关系。典型的在线社区网站有 Yahoo 多功能社区网、斯坦福大学学生在线社区、专业人士推荐网站 Linked-in 以及微软公司的博客网站 MSN Space 等。

（2）在线交友。近年来在互联网中出现了大量的社交网络服务站点，这些站点依据六度分隔理论，通过朋友的介绍来结交新的朋友，不断扩大自己的交友范围，扩展自己的人脉。在线交友网站中，一般把用户作为节点，用户的好友关系作为边，当一个注册用户通过邀请与另一个用户建立了好友关系，则相当于在两个节点之间建立了一条边。这种网络社会化使互联网应用从传统的“人机对话”模式逐渐转变为“人人对话”模式。典型的在线交友网站有 Facebook、人人网、MySpace 等。

（3）在线网络媒体。在互联网中还有一种称为在线网络媒体的网站，这类网站主要提供发布、分享和检索新闻、图片、音频、视频等媒体资源的功能，具有即时性、海量性、全球性、交互性等特点，如全球最大的视频分享网站 YouTube、相片分享网站 Flickr 等。另外，在线网络媒体不仅仅是一个资源分享网站，还是一个以资源分享为纽带的用户关系网站。这些在线网络媒体正在逐渐改变网络用户交互与信息交流的方式。

5．在线社交网络问题

在线社交网络是围绕着用户来建立和组织的，可以利用在线社交网络建立的朋友关系进行广告宣传、产品推介、信息交流、活动联络等，在促进人们的社会交往和信息交流中起到积极的作用，同时也会被不法的组织和人员利用进行非法联络和谣言传播，例如，在国内外发生的暴力恐怖事件中，恐怖组织利用在线社交网络进行谣言散布和非法联络；在国外发生的“颜色革命”中，不同程度地利用了在线社交网络进行集会联络和信息传播。可见，在线社交网络是一把双刃剑，由在线社交网络引发的网络安全问题将给社会和谐稳定以及国家安全带来极大的威胁。

在线社交网络的广泛应用，推动了国内外关于在线社交网络的研究工作，研究人员从不同的角度对在线社交网络进行了研究，内容涉及在线社交网络的基本特性、网络结构、信息传播、用户关系、连接强度等方面，通过建立相应的数学模型，对在线社交网络特性进行分析，找出其中的信息传播特性和内在规律，为优化在线社交网络结构、改善在线社交网站服务、实施在线社交网站监管等提供科学依据和解决方案。

1.4　微博网络概论

微博（Microblogging）网站是一种集成化、开放式的互联网社交服务平台，用户通过 140 字以内的微博发布信息，实现即时分享。此外，用户可以根据自己的兴趣偏好，选择关注其他用户，构建自己的关注网络。

2006 年 3 月，博客技术的创始人威廉姆斯所创建的互联网公司Obvious开发并推出了 Twitter 网站。Twitter的出现把人们带入了一个全新的互联网时代，即微博时代。关于名字 Twitter 的来历，其英文原意为鸟儿的叽叽喳喳声，创始人认为鸟儿的叫声具有短、频、快的特点，符合该网站的内涵，因此选择了 Twitter 作为网站的名称。在最初阶段，Twitter 只提供向好友的手机发送文本信息的服务，后来逐渐增加了一些新的服务，比如，用户可以通过SMS、电邮、Twitter网站或 Twitter 客户端软件（如 Twitterrific）接收和发送信息，现在的 Twitter网站已发展成一个集社交网络和微博为一体的综合社交服务网站。

此后，国内外出现了大量类似 Twitter 的网站，国外的有Plurk、Jaiku等，国内的有饭否、做啥、叽歪、嘀咕、贫嘴、同学网、腾讯滔滔、9911 等，其中，饭否影响力较大，2009 年上半年，饭否的用户从年初的 30 万激增到 100 万，随着众多文化名人的加入以及国内众多知名媒体开辟饭否官方账号，饭否一度成为中国微博的标杆。后来，国内的四大门户网站均开设了微博网站，微博用户数量迅猛地增长。尽管近几年受到微信等即时通信工具的冲击，但微博的网民数仍然是比较庞大的。

Twitter 网站是世界上率先推出的微博平台，以崭新的信息交流方式在世界上引起极大的反响，成为全球影响力最大的微博平台。新浪微博是国内最大的微博平台，其注册用户数超过 5 亿人，日活跃用户数达到 4620 多万人。

微博打破了传统媒体单一的舆论主导权，给大众提供了一个自由发表意见并与他人分享的平台，在一定程度上保证了公众的话语权。因此，微博极大地解放了公众话语权，促使了公众话语权的回归，开创了一个平民化的信息传播模式。

微博网络作为一种特殊的社交网络，用户不但可以有选择地连接感兴趣的用户，关注其信息，而且也可以被其他用户相互连接，交流信息，具有社交网络和媒体网络的双重特性，一些学者认为微博网络是一个社交媒体网络。

微博作为新兴的社交媒体，越来越受到重视。国外的许多政治人物都将微博作为推广政见的工具。例如，美国总统特朗普在竞选期间通过 Twitter 与选民进行交流互动，赢得了更多选民的支持。国外的一些政府部门、新闻机构等都开通了 Twitter 账号，作为与民众沟通交流、获取信息的手段。在国内，自 2009 年云南省政府新闻办开设了国内第一家政府微博“微博云南”后，全国各地的政府部门都陆续开通了政务微博，实时发布消息，与民众互动。国内的 CCTV、人民日报、新华通讯社等主流媒体也都在新浪微博平台上开通了官方微博。

由于微博网络在当今社会信息传播中发挥越来越重要的作用，同时微博网络作为一种特殊的社交网络，存在着与社交网络相类似的网络安全问题，如非法联络、传播谣言、煽动闹事等，容易引发社会群体事件，给社会和谐稳定以及国家安全带来极大的威胁。

微博转发是微博网络提供的一种信息传播机制，用户可以将关注者发布的微博转发到自身平台上，然后分享给粉丝。通过这种信息传播机制，使得微博能够在更大范围内传播和分享。可见，用户转发行为是推动微博信息传播的重要因素。

微博网络的广泛应用也引起了国内外学术界的关注，研究人员对微博网络的基本特性、网络结构、信息传播、用户行为等方面进行了研究，通过建立相应的数学模型，对微博用户转发行为特性进行分析，找出其中的内在规律，不仅可以为网络舆情监测、突发事件预测等提供科学依据，还可以为商家分析用户购买喜好、推荐商品以及精准投放广告等提供参考和帮助。

1.5 网络论坛概论

网络论坛是一种为用户提供信息交流平台的网络应用系统，网络论坛也称为电子公告板BBS（Bulletin Board System），最早是用来发布股市价格等信息的，当时的 BBS 功能比较简单，连文件传输功能都没有。随着计算机技术和网络技术的发展，以及网络信息交流需求的驱动，BBS 不断发展壮大，现在的网络论坛几乎涵盖了社会生活的方方面面，每个用户都可以找到自己感兴趣或者需要了解的专题性论坛。综合性门户网站和功能性专题网站等各类网站也都开设了自己的论坛，以促进网民之间的交流，增强网民的互动性。

网络论坛属于传统的网络信息交流平台，随着社交网络、微博网络等新型网络信息交流平台的广泛应用，网络论坛的用户数量有所下降，尽管其网民数和使用率不如微博、在线社交网络高，但网络论坛所具有的多元化、开放性、匿名性及互动性，仍然是广大网民发表言论、获取信息的重要网络平台，用户数量还是比较庞大的。

1．网络论坛类型

网络论坛总体上可分为综合类论坛和专题类论坛两大类。综合类论坛包含的信息比较丰富和广泛，能够吸引很多的网民来到论坛，但往往广而不精。

专题类论坛专注于特定的专题，例如军事类论坛、情感倾诉类论坛、电脑爱好者论坛、动漫论坛等，能够吸引志同道合的人参与交流讨论，有利于信息的分类整合和搜集。专题类论坛能够在一个单独的领域里进行版块的划分和设置，使得专题更加细化，取得更好的效果。

网络论坛种类有很多，如教学类论坛、推广类论坛、地方性论坛和交流性论坛等。

- 教学类论坛主要提供教学交流和知识传播的场所，通过在论坛上浏览帖子和发布帖子，能够与他人在网上进行知识学习和教学交流。

- 推广类论坛主要用于企业及产品的宣传推广，也是一种广告形式。这样的论坛很难有吸引人的性质，往往寿命很短，论坛中的会员基本是受雇佣的人员。
- 地方性论坛是论坛中娱乐性与互动性最强的论坛之一，不论是大型论坛中的地方站，还是专业类的地方论坛，都有大量的网民参与其中。地方性论坛能够拉近人与人的沟通和交流，具有地方性特色，因此受到网民的欢迎。
- 交流性论坛的重点在于论坛成员之间的交流和互动，其内容比较丰富多样，有供求信息、交友信息、线上线下活动信息以及新闻等，这类论坛是将来论坛发展的趋势。

2．网络论坛舆论

网络论坛以开放性、匿名性及互动性为特色，为网民提供了发表言论、获取信息的网络信息交流平台。在网络论坛中，网民就某个主题通过发帖、观看和回帖进行信息交流和互动，在信息交流过程中，某些话题的帖子受到网民的高度关注，点击量和回帖数非常大，形成较大的影响力，这种帖子称为热帖，热帖在观点传播和舆论形成过程中起到重要的推动作用。

可见，网民通过发帖和回帖发表意见，参与观点传播和舆论形成，对于推进社会进步和政治民主起到了积极的作用，成为网络舆情的主要来源。

所谓舆论是指公众对社会某些事务或现象的一致意见表达。网络舆论也具有舆论的本质属性，公众或网民以网络为平台，就某些事务或现象发表意见，表达观点，可以看作是一种特殊的舆论形式。而网络舆情是指网民就某些社会问题或公共事务表达不同看法的网络舆论，反映了公众对现实生活中的某些热点、焦点问题所持的具有较强影响力和倾向性的言论和观点。

在网络舆论形成过程中，意见领袖起到了积极的推动作用。统计数据显示，网络中的大部分用户不经常参与信息的制造与传播，他们做出的决定往往跟随意见领袖。通过意见领袖发表引导性意见来影响所在网络用户而非直接说服他们，可以有效地触发整个网络舆论的影响力，能够有效地推动信息的传播，提高广告效应。同时，网络论坛也是一把双刃剑，它所具有的开放性和匿名性等特点，容易被别有用心的组织和人员所利用，传播虚假消息和谣言，对人们的社会生活和意识形态造成负面的影响。

3．网络水军问题

截至 2015 年 12 月，我国的网民已达到 6.88 亿人，很多网民将互联网视为了解社情民意、揭露社会弊端、开展社会监督的窗口。2009 年以来，南京“天价香烟”事件、河南民工“开胸验肺”事件、云南晋宁县“躲猫猫”事件等热点事件，均由网络舆论率先关注，继而引发媒体报道。据中国社科院发布的《蓝皮书》透露，在 2009 年 77 件影响力较大的社会热点事件中，由网络爆料而引发公众关注的有 23 件，约占全部事件的 30%。可

见，互联网已成为新闻舆论监督的重要平台，特别是以开放性、匿名性及互动性为特点的网络论坛成为网络舆论的主要来源。

然而，随着网络舆论对社会和公众影响的不断增大，出现了以网络炒作为营生的网络公关公司、网络推手、网络水军等。网络公关公司受托于客户，为了在网上炒作某个话题或人物或产品来达到宣传、推销或者诋毁他人或产品的目的，雇佣了大量的网络水军，在网络推手的组织下，以各种手法和名目在互联网的各大网络论坛上短时期内大量地发帖和回帖，炮制网络热点事件，捧红各色人物，形成虚假的网络舆情。例如，在央视感动中国2010 年度人物评选中就遭遇网络水军的密集刷票，引起社会各界高度关注；通过网络炒作，使“奥巴马女郎”“兽兽门”“阎凤娇裸照门”“凤姐”“犀利哥”等原本无名人物在一夜之间名扬网络；在网络上被传得沸沸扬扬的“王老吉”添加门、“360”曝黑门、“康师傅”水源门、“伊利”牛奶门等事件都是通过网络炒家人为炒作出来的。

网络公关公司、网络推手、网络水军等形成了灰色利益链，他们在实现客户目标的同时也获得自身利益。据公安部门调查，当前国内一些大的网络论坛，有 50%左右的帖子是人为炒作推出来的。所谓“热门帖”“精华帖”等，很少是网民自发点击、回帖形成的，背后几乎都有网络炒家在积极推动，都是由网络水军实施的，这种虚假网络舆情被称为网络灌水现象。

网络水军及其网络灌水问题具有很大的危害性，在网络舆情中存在歪曲失真信息泛滥、网民群体极化倾向严重、境内外不法分子恶意操纵、国外敌对势力渗透性入侵等安全隐患，产生错误的舆论导向，危及政府的公信力，引发社会群体性事件等问题。对于网络水军所产生的负面影响，已引起新闻媒体和国家有关部门的关注，央视等新闻媒体多次对网络水军问题进行采访报道和深度分析；国家互联网管理部门制定了加强互联网管理的有关规定，并依法惩戒了利用互联网进行造谣惑众、恶意炒作的不法网民，包括“秦火火”“立二拆四”等网络名人。

网络水军及虚假网络舆情问题引起了社会和学术界的极大关注，研究人员通过建立相应的数学模型，对网络论坛的信息传播特性、网络舆情检测、意见领袖发现、网络水军识别等问题进行了研究，找出其中的内在规律，为快速检测网络舆情、识别虚假网络舆情、抑制网络谣言传播提供科学依据和解决方案。

第2章

网络建模基本理论

2.1 引言

本书主要运用复杂网络理论、传播动力学等方法对 P2P 网络、社交网络、微博网络、网络论坛等网络信息交流平台的信息传播特性和机理进行建模分析。由于在以下各章中均涉及复杂网络理论、传播动力学等网络建模基本理论，为了叙述方便，本章统一对网络建模基本理论进行简单的介绍。

2.2 网络的图表示方法

图论是数学的一个分支，它以图为研究对象。图论中的图是由若干给定的点及连接两点的线所构成的图形，这种图形通常用来描述某些事物之间的某种特定关系，用点代表事物，用连接两点的线表示相应两个事物间具有某种关系。

定义 2-1：网络可表示为点集 V 和边集 E 组成的图 G，记作 $G=(V,E)$，且

（1）$V=\{v_1,v_2,\cdots,v_{N_V}\}$ 是顶点的集合，其中的元素 v_i 表示网络中的具体节点，$1\leqslant i\leqslant N_V$，$N_V$ 表示网络中节点的数目。图 G 中所有节点的集合可用 $V(G)$ 表示；

（2）$E=\{e_1,e_2,\cdots,e_{N_E}\}$ 是边的集合，其中的元素 $e_j=(v_{j_1},v_{j_2})$ 表示网络中节点 v_{j_1} 与 v_{j_2} 之间的连接，一般来讲，v_{j_1} 与 v_{j_2} 不是同一个节点，$1\leqslant j\leqslant N_E$，$N_E$ 表示网络中边的数目。图 G 中所有边的集合可用 $E(G)$ 表示。

定义 2-2：当图 G 中任意两个节点对 (v_{j_1},v_{j_2}) 和 (v_{j_2},v_{j_1}) 表示同一条边时，则 G 称为无向图；否则，图中的边称为有向边，图 G 称为有向图，有向图也可以用 G_d 表示。在后续的论述中，如果不做特殊说明，图 G 表示无向图。

定义 2-3：在图 $G=(V,E)$ 中，如果图 G 中的每一条边 $e_j=(v_{j_1},v_{j_2})$，都有一个权重 w_j，则图 G 称为赋权图；否则，图 G 称为无权图。

定义 2-4：对图 G 和图 H 来说，当 $V(H)\subseteq V(G)$，且 $E(H)\subset E(G)$ 时，图 H 是图 G 的子图，记作 $H\subseteq G$。当 $H\neq G$ 时，图 H 是图 G 的真子图，记作 $H\subset G$。

定义 2-5：度（degree）是节点属性中的重要概念，节点 v_i 的度是指与该节点相连接的其他节点数量，用 $\deg(v_i)$ 或 k_i 表示。在有向图 G_{d} 中，节点的度可分为入度和出度：节点的入度是指从其他节点指向该节点的边的数目，用 $\deg_{\mathrm{in}}(v_i)$ 或 k_i^{in} 表示；节点的出度是指从该节点指向其他节点的边的数目，用 $\deg_{\mathrm{out}}(v_i)$ 或 k_i^{out} 表示，且 $\deg(v_i)=\deg_{\mathrm{in}}(v_i)+\deg_{\mathrm{out}}(v_i)$。图 G 的度 $\deg(G)$ 为图中所有节点的最大度值，即：

$$\deg(G)=\max_{i=1}^{N_V}(\deg(v_i)) \tag{2-1}$$

如果图 G 中所有节点的度都是常数 k，那么称图 G 为规则图，如果 k 的值为 N_V-1，那么任意一个节点与任意其他节点都有边相连，则称图 G 为完全图。

定义 2-6：图 G 中 2 个节点 v_i 和 v_j 的距离 d_{ij} 定义为连接这两个节点的最短路径上的边数。如果两个节点不可达，那么它们之间的距离为无穷大（∞）。

定义 2-7：图 G 的直径 D_{G} 定义为图中任意节点对距离的最大值，即：

$$D_{\mathrm{G}}=\max_{i,j=1,i\neq j}^{N_V}(d_{ij}) \tag{2-2}$$

定义 2-8：从一个节点出发沿着图 G 中的边所能到达的全部节点集合，称为图 G 的一个联通子图。对图 G 而言，如果从一个节点出发沿图中的边能够到达图中的任何节点，则称图 G 为连通图。

2.3 复杂网络基本理论

2.3.1 复杂网络基本概念

网络与图的最早研究起源于解决“哥尼斯堡七桥”问题，随后逐步发展成为系统化的科学理论。随着 Erdős 和 Rényi 提出随机网络（Erdős-Rényi，ER）模型[1]，面向真实网络的建模及理论研究取得了很大的进展。但是随着计算机处理能力的提高，网络研究由几百个节点的小网络，转向了规模更大、结构更复杂的网络系统，人们发现随机网络模型在处理大规模复杂网络时变得无能为力，并且很多真实网络的特性无法用随机网络模型来解释。这些问题使得对复杂网络的科学理解成为网络理论研究中一个极具挑战性的课题。

1998 年，Watts 和 Strogatz 引入了小世界（Small World）网络模型[2]，该模型以小概率改变规则网络中边的连接方式，构造出介于规则网络和随机网络之间的网络，该网络既

具有高聚类特性，又具有较小平均路径长度。1999 年，Barabási 和 Albert 通过在互联网的随机访问发现互联网的度分布符合“胖尾”幂律分布[3]，并指出许多实际复杂网络的度分布具有幂律形式。由于幂律分布没有明显的特征长度，该类网络又被称为无标度（Scale-Free）网络。随着这些开创性研究的进展，复杂网络的科学探索发生了重要转变，开辟了复杂网络研究的新方向。

大部分现实网络无论从规模还是网络结构来看，都是复杂网络。例如，代谢网络[4]、蛋白质网络[5]、神经网络[6]、电影演员关系网络[7]、科学家合作网络[8, 19]、电子邮件网络[10, 11]、电力网络[12]、互联网[13]、Web 网络[14]、P2P 网络[15]等。虽然复杂网络已成为研究热点，然而目前人们还没有给出它的精确定义。比较公认的复杂网络具有三个特征：小世界效应、自由标度性和高聚类性。

目前，复杂网络研究涉及范围比较广泛，在国际一流刊物上发表了大量的文章，反映了复杂网络已经成为国际学术界的研究热点。总体上，复杂网络的研究内容可以归纳为以下几类。

1．网络拓扑特性分析

网络拓扑特性分析是研究复杂网络的最基本手段，目的是发现复杂网络的一些统计特性，例如连接度与度分布、平均路径长度与聚类系数、拓扑层次化等，并研究相关特性的有效评价方法，试图认识和掌握各种内在的规律。

2．复杂网络建模研究

图论中提出的经典模型已经被证明与实际网络相差较远，必须发展新的网络模型来模拟网络的生长过程以及重现那些在实际网络中观察到的结构属性。根据对各种实际复杂网络数据的分析结果，概括出它们的共有特性，再结合对实际网络形成机制的理解和解释，通过生成算法构建符合真实网络统计特性的网络演化模型，模仿真实网络行为，再现真实网络几何特性。

3．复杂网络动力学研究

每个复杂网络都是一个复杂的动力系统，由节点所代表的动力学单元相互作用构成。复杂网络动力学研究主要包括：网络结构如何影响动态属性，如鲁棒性和同步能力等；混沌动力系统在网络上的同步性；网络拥塞及信息在复杂网络上的传播；小世界网络的自组织临界现象；复杂网络控制问题等。理解了网络上各种复杂行为的内部机制，有利于更加有效地实施控制策略和资源配置。

4．复杂网络应用研究

尽管复杂网络理论还在完善中，但复杂网络已经开始应用到各个学科领域中，主要包

括：根据复杂网络模型挖掘与功能相关的深层内容；应用复杂网络鲁棒性研究成果进行网络设计；应用传播动力学理论研究流行病传播过程；将小世界网络思想应用于人工神经网络，可以减少神经网络的学习时间和学习误差；将复杂网络理论应用于 Hopfield 神经网络，可以改变 Hopfield 神经网络的联想记忆功能。随着复杂网络研究的不断发展，将会有越来越多的问题通过复杂网络理论来解决。

2.3.2 复杂网络拓扑特征

复杂网络的拓扑特征往往决定了该系统所具有的功能特性。因此，人们对复杂网络的研究，更多的是立足于对其拓扑特征的研究。虽然真实的复杂网络在网络规模、节点属性、承载功能等方面差异较大，但大量研究表明，这些复杂网络普遍存在一些共同的拓扑特征。研究人员提出了许多刻画复杂网络拓扑特征的概念，这些概念在研究中起到了至关重要的作用，下面对主要概念进行介绍。

1．平均路径长度与全局效率

平均路径长度 $<d>$ 定义为网络中任意两个节点之间距离的平均值，即：

$$<d>=\frac{1}{\frac{1}{2}N_V(N_V-1)}\sum_{i,j=1,i\neq j}^{N_V}d_{ij} \tag{2-3}$$

式中，N_V 为网络节点数，在实际应用中，N_V 的数量级一般很大。如果是非连通网络，部分节点对之间没有连通路径，它们的距离为无穷大，$<d>$ 的计算结果会变为无穷大。为了解决这个问题，一方面，定义 $<d>$ 为所有存在连通路径节点对的平均最短路径长度，将没有连通路径的节点对排除在外；另一方面，使用全局效率 E_G 来代替 $<d>$ 描述网络的功能特性。全局效率 E_G 的定义为：

$$E_G=<d>^{-1}=\frac{1}{\frac{1}{2}N_V(N_V-1)}\sum_{i,j=1,i\neq j}^{N_V}\frac{1}{d_{ij}} \tag{2-4}$$

式中，$1/d_{ij}$ 表示节点对之间的传输效率，用来描述网络传递信息的能力，避免了定义 d_{ij} 时出现无穷大的情形，对于不连通路径节点对来说，$d_{ij}=\infty$，$1/d_{ij}=0$。全局效率 E_G 也称为最短路径长度的调和平均数。

尽管现实世界的许多复杂网络节点数巨大，但是网络的 $<d>$ 都相对较小，即使是稀疏网络也是如此。Watts 和 Strogatz 指出，$<d>$ 与网络规模存在一定的关系，当网络规模增加时，$<d>$ 通常也将随之增大。若 $<d>$ 的增加是 $\ln N_V$ 的阶数，则认为这种网络的平均路径比较小，称为“小世界”现象。如电影演员合作网络的 $<d>$ 为 3.65，Web 网络的 $<d>$ 为 3.11。

2．度分布

复杂网络中所有节点 v_i 的度 $\deg(v_i)$ 的平均值称为网络平均度 $<k>$，即：

$$<k>=\frac{1}{N_V}\sum_{i=1}^{N_V}\deg(v_i) \tag{2-5}$$

复杂网络的度分布使用节点度的概率分布函数 $P(k)$ 来描述，表示随机选定一个节点，其度值恰好为 k 的概率，也就是节点有 k 条边连接的概率，即：

$$P(k)=\frac{N_k}{N_V} \tag{2-6}$$

式中，$1\leqslant k\leqslant N_V-1$，$N_k$ 表示网络中度数为 k 的节点数。另一种描述度统计特性的方法是累积度分布 P_k，表示节点度数大于或等于 k 的概率，即：

$$P_k=\sum_{k'=k}^{N_V-1}P(k') \tag{2-7}$$

采用 P_k 表示分布有两个好处：一是保持了单点突变现象，二是减弱噪声干扰的影响。

$P(k)$ 的 n 阶距是另外一种刻画复杂网络节点度分布的物理量，定义为：

$$<k^n>=\sum_{k=1}^{N_V-1}k^nP(k) \tag{2-8}$$

式中，一阶矩对应网络平均度 $<k>$，二阶矩 $<k^2>$ 刻画了度分布波动大小。

3．度相关性

度分布反映了无关联网络的统计特性，但许多真实复杂网络的节点度值之间存在一定关联性。度相关性主要考查节点度之间的关联，如果度大的节点倾向于和度大的节点连接，则复杂网络是正相关的；反之，复杂网络是负相关的。

度关联性有两种表示方法，一种方法是直接使用联合度分布函数 $P(k,k')$，表示任意一条边的两个端点的度分别为 k 和 k' 的概率，对于无向网络来说，$P(k,k')=P(k',k)$；另一种方法是使用条件概率 $P(k'|k)$ 描述节点度之间的关联，$P(k'|k)$ 表示任意一条边的起点度为 k，终点度为 k' 的概率，它满足归一化条件和节点度的平衡条件[16,17]，即：

$$\begin{cases}P(k'|k)=\dfrac{<k>P(k,k')}{kP(k)}\\ \displaystyle\sum_{k'=1}^{N_V-1}P(k'|k)=1\\ k'P(k|k')P(k')=kP(k'|k)P(k)\end{cases} \tag{2-9}$$

形式上，$P(k,k')$ 和 $P(k'|k)$ 刻画了节点的度关联性。由于网络大小是有限的，直接计算它们比较困难，而且会产生很大的噪声。为了更加方便地判断网络度相关性，Newman 给出了一种更加简便的计算方法[18]，只需计算节点度的 Pearson 相关系数 r 即可，即：

$$r=\frac{\frac{1}{N_E}\sum_{i=1}^{N_E}k_{i1}k_{i2}-\left[\frac{1}{N_E}\sum_{i=1}^{N_E}\frac{1}{2}(k_{i1}+k_{i2})\right]^2}{\frac{1}{N_E}\sum_{i=1}^{N_E}\frac{1}{2}(k_{i1}^2k_{i2}^2)-\left[\frac{1}{N_E}\sum_{i=1}^{N_E}\frac{1}{2}(k_{i1}+k_{i2})\right]^2} \tag{2-10}$$

式中，N_E 表示复杂网络的总边数，$1\leqslant i\leqslant N_E$，$k_{i1}$ 和 k_{i2} 表示第 i 条边的两个顶点 v_{i1} 和 v_{i2} 的度。r 的取值范围为 $-1\leqslant r\leqslant 1$，当 $r>0$ 时，网络是正相关的；当 $r<0$ 时，网络是负相关的；当 $r=0$ 时，网络是不相关的。Newman 计算了一些复杂网络的 r，发现社交网络是正相关的，技术网络和生物网络是负相关的。

4．聚类系数

社交网络的一个普遍特点是小聚类现象，例如在朋友网络中，很容易发现你朋友的朋友也是你的朋友，这种特征称为聚类特征。为了刻画这种网络集团化程度，使用聚类系数来衡量复杂网络中节点之间连接的紧密程度，它反映了网络中三角形结构密度，网络中的三角形分布越密集，说明网络聚类性越强。聚类系数可以针对单个节点度量，也可以针对网络整体度量。

在复杂网络中，节点 v_i 的度为 $\deg(v_i)$，也可以使用 k_i 表示，表示有 k_i 条边将它和其他节点直接相连，相应地，这 k_i 个节点称为节点 v_i 的最近邻居，在这 k_i 个邻居节点之间最多可能存在 $k_i(k_i-1)/2$ 条边。因此，定义节点 v_i 的聚类系数 c_i 为在 k_i 个邻居节点之间实际存在的边与可能存在的边之比，即：

$$c_i=\frac{E(\Gamma(v_i))}{\frac{1}{2}k_i(k_i-1)} \tag{2-11}$$

式中，k_i 为节点 v_i 的度，$\Gamma(v_i)$ 为节点 v_i 的邻居节点所形成的子图，$E(\Gamma(v_i))$ 表示 $\Gamma(v_i)$ 中的边数，也就是节点 v_i 的 k_i 个邻居节点之间实际存在的边数。复杂网络的聚类系数 C_G 定义为网络中所有节点的聚类系数平均值，即：

$$C_G=\frac{1}{N_V}\sum_{i=1}^{N_V}c_i \tag{2-12}$$

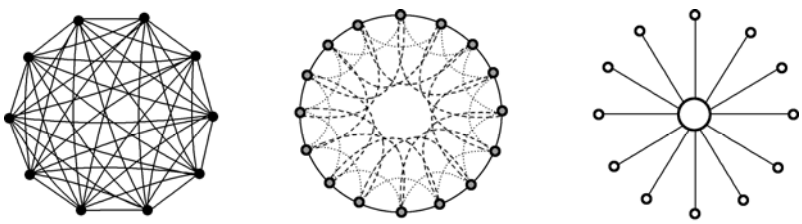
从定义可以看出，$0\leqslant c_i\leqslant 1$，$0\leqslant C_G\leqslant 1$。只有当网络是全局耦合网络，任意两个节点都直接相连时，$C_G=1$。对于一个含有 N_V 个节点的完全随机网络，当 N_V 很大时，

$C_G = O(N_V^{-1})$。而许多实际大规模复杂网络都有明显的聚类效应，它们的聚类系数尽管远小于 1，但却比 $C_G = O(N_V^{-1})$ 要大得多。这意味着这类网络并不是完全随机的，而是在某种程度上具有类似于社会关系网络中“物以类聚、人以群分”的特征。复杂网络研究中，微观上的强聚类现象、小世界效应和连接度的幂律分布三个特征成为衡量复杂网络的三大标志性特征。

2.4　经典网络模型

2.4.1　规则网络模型

如果节点之间按照确定的规则连线，得到的网络称为规则网络。最常见规则网络包括：全局耦合网络、最近邻耦合网络、星形网络，如图 2-1 所示。



（a）全局耦合网络　（b）最近邻耦合网络　（c）星形网络

图 2-1　常见规则网络

全局耦合网络中任意两点都有边直接相连，在具有相同节点数的网络中，全局耦合网络具有最小的平均路径长度 $< d >=1$ 和最大的聚类系数 $C_G = 1$。N_V 个节点的全局耦合网络具有 $N_V(N_V - 1)/2$ 条边，然而大多数实际网络都是稀疏的，它们的边数一般是 $O(N_V)$。最近邻耦合网络中 N_V 个节点围成一个环，每个节点都与它左右各 $K_{\mathrm{nc}}/2$ 个邻居节点相连，K_{nc} 为偶数。最近邻耦合网络的聚类系数为：

$$C_G = \frac{3(K_{\mathrm{nc}} - 2)}{4(K_{\mathrm{nc}} - 1)} \tag{2-13}$$

当 K_{nc} 较大时，$C_G \approx 3/4$。最近邻耦合网络的平均路径长度为：

$$< d >\approx \frac{N_V}{2K_{\mathrm{nc}}}\bigg|_{N_V \to \infty} \to \infty \tag{2-14}$$

星形网络有一个中心点，与其他 $N_V - 1$ 个节点相连，而这 $N_V - 1$ 个节点之间没有任何边相连。因此，星形网络的聚类系数为 0，平均路径长度为：

$$< d >\approx 2-\frac{2(N_V-1)}{N_V(N_V-1)}\Bigg|_{N_V\to\infty}\to 2 \tag{2-15}$$

规则网络具有如下重要的特性。

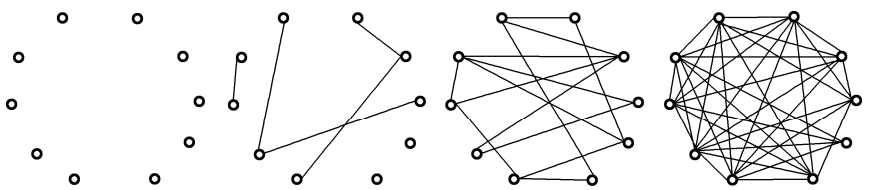
（1）具有均匀的度分布；

（2）聚类系数几乎与网络大小无关，而且比随机网络大得多；

（3）与随机网络相比，规则网络的平均路径长度要大得多，而且随网络规模的增长而不断增长，当网络规模趋于无穷时，其平均路径长度也趋于无穷。

2.4.2 随机网络模型

随机网络模型是以相同概率 p 连接随机选定节点对，若节点总数为 N_V，则网络边数为 $pN_V(N_V-1)/2$。使用这种方法生成的所有网络群体可以使用 $G_{N_V,p}$ 表示。演化过程如图 2-2 所示。



（a）$p=0$　（b）$p=0.1$　（c）$p=0.3$　（d）$p=0.8$

图 2-2　随机网络演化示意图

随机网络的基本性质归纳如下。

1．涌现性质

当随机网络节点总数 $N_V\to\infty$ 时，随机网络的结构和性质都随概率 p 而变化，它的很多重要性质都是在某个临界概率 p_c 处突然涌现出来的。

2．平均度

在随机网络中，任一节点都以概率 p 与其他 N_V-1 个节点相连，所以其平均度为 $< k >= p(N_V-1)\approx pN_V$。

3．度分布

在连接概率为 p 的随机网络模型中，节点 v_i 的度为 k 的概率符合参数为 N_V-1 和 p 的二项式分布，即 $P(k)=C_{N_v-1}^{k}p^k(1-p)^{N_v-1-k}$。节点 v_i 引出 k 条边与 k 个节点相连的概率为 p^k，与 N_V-1-k 个节点不相连的概率为 $(1-p)^{N_v-1-k}$，共有 $C_{N_v-1}^{k}$ 种方式选择 k 个节点。当网络规模 $N_V\to\infty$ 时，度分布为：

$$P(k)=C_{N_V-1}^{k}p^{k}(1-p)^{N_V-1-k}\approx e^{-<k>}\frac{(<k>)^{k}}{k!} \tag{2-16}$$

这种大多数节点度值都在$<k>$附近，没有度值特别大的节点。度的分布符合泊松分布，表明随机网络是一种均匀网络，节点之间的连接是等概率的。

4．平均路径长度

在随机网络中，与节点v_i距离为 d 的节点数为$<k>^{d}$，包含所有节点的 d 应满足$<k>^{d}=N_V$，因此随机网路的平均路径长度$<d>$为：

$$<d>=\frac{\ln N_V}{\ln <k>} \tag{2-17}$$

公式（2-17）表明，随机网络的平均路径长度对节点总数的增加呈对数增长，规模很大的随机网络具有较短的平均路径长度。

5．聚类系数

随机网络中任何两个节点之间的连接都是等概率的，因此，聚类系数为：

$$C_G\approx p=\left.\frac{<k>}{N_V}\right|_{N_{V\to\infty}}\to 0 \tag{2-18}$$

公式（2-18）表明，当$N_V\to\infty$时，随机网络的聚类系数趋近于 0，没有聚类特性。

随机网络具有度分布服从泊松分布、较小平均路径长度和较小聚类系数等性质。该模型被大多数人所接受，很多网络拓扑结构采用该模型描述。但是，随着计算机处理能力的增强，研究人员发现大量现实网络并不是完全随机网络，它们具有较大的聚类系数，而随机网络聚类系数很小。虽然人们对随机网络模型进行了多角度扩展，但是这些扩展并没有从本质上解决刻画真实网络时存在的问题。

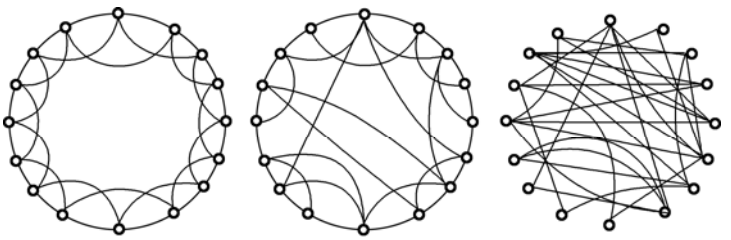
2.4.3　小世界网络模型

随机网络虽然具有较小平均路径长度，但没有高聚类特性，难以刻画现实复杂网络的小世界特性。因此，人们提出了小世界网络模型，作为从规则网络向随机网络的过渡，其主要模型有：Watts 和 Strogatz 提出的 WS（Watts-Strogatz）模型[2]、Newman 和 Watts 提出的 NW（Newman- Watts）模型[19]等。WS 模型的构造算法如下。

（1）构造一个由N_V个节点组成的最近邻耦合网络，网络中的每个节点都与其左右相邻的$K_{\mathrm{nc}}/2$个节点相连，K_{nc}为偶数；

（2）以固定概率 p 随机重连网络中的每条边，即将边的一端保持不变，而另一端为随机选择的一个节点。并且网络中任意两个不同节点之间至多只能有一条边，并且每一个节点都不能有边与自身相连。

WS 模型通过以概率 p 重新连接网络中已经存在的边，构造出一种介于规则网络和随机网络之间的网络。规则网络模型和随机网络模型分别是 WS 模型在 $p=0$ 和 $p=1$ 时的特例。使用不同概率构造的网络如图 2-3 所示。



（a）$p=0$　　（b）$p=0.2$　　（c）$p=1$

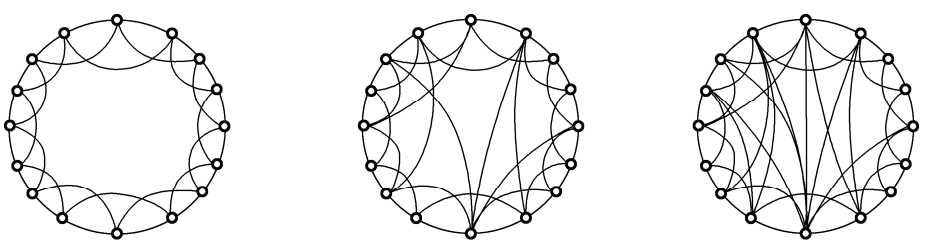
图 2-3　WS 网络模型

NW 模型通过“随机化加边”来取代 WS 模型中的“随机化重连”，该模型的构造算法如下：

（1）构造一个由 N_V 个节点组成的最近邻耦合网络，网络中的每个节点都与其左右相邻的 $K_{\mathrm{nc}}/2$ 个节点相连，K_{nc} 为偶数；

（2）以固定概率 p 随机地选择一对节点进行连接，并且网络中任意两个不同节点之间至多只能有一条边，并且每一个节点都不能有边与自身相连。

NW 模型构造出一种介于规则网络和全耦合网络之间的复杂网络，当 $p=0$ 时，构造网络为规则网络，当 $p=1$ 时，构造网络为全耦合网络。使用不同概率构造的网络如图 2-4 所示。



（a）$p=0$　　（b）$p=0.05$　　（b）$p=0.1$

图 2-4　NW 网络模型

NW 模型相对于 WS 模型来说，算法简单，容易实现，同时避免了 WS 模型中由于重连而可能产生孤立点情况的发生。当 p 足够小和 N_V 足够大时，NW 模型本质上等同于 WS 模型。它们的拓扑特性主要表现为以下几点。

1．度分布

当 $0<p<1$ 时，WS 模型的节点度分布为[20]：

$$\begin{cases} P(k)=\displaystyle\sum_{i=0}^{\min\left(k-\frac{K_{\mathrm{nc}}}{2},\frac{K_{\mathrm{nc}}}{2}\right)}\binom{\frac{K_{\mathrm{nc}}}{2}}{i}(1-p)^{i}p^{\frac{K_{\mathrm{nc}}}{2}-i}\dfrac{\left(p\frac{K_{\mathrm{nc}}}{2}\right)^{k-\frac{K_{\mathrm{nc}}}{2}-i}}{\left(k-\frac{K_{\mathrm{nc}}}{2}-i\right)!}\mathrm{e}^{-p\frac{K_{\mathrm{nc}}}{2}} & k\geqslant\dfrac{K_{\mathrm{nc}}}{2} \\ P(k)=0 & k<\dfrac{K_{\mathrm{nc}}}{2} \end{cases} \tag{2-19}$$

当 $0<p<1$ 时，NW 模型的节点度分布为[21]：

$$\begin{cases} P(k)=\binom{N_V}{k-K_{\mathrm{nc}}}\left(\dfrac{K_{\mathrm{nc}}p}{N_V}\right)^{k-K_{\mathrm{nc}}}\left(1-\dfrac{K_{\mathrm{nc}}p}{N_V}\right)^{N_V-k+K_{\mathrm{nc}}} & k\geqslant K_{\mathrm{nc}} \\ P(k)=0 & k<K_{\mathrm{nc}} \end{cases} \tag{2-20}$$

2．平均路径长度

NW 模型的平均路径长度为：

$$<d>=\frac{2N_V}{K_{\mathrm{nc}}}f\left(\frac{pK_{\mathrm{nc}}N_V}{2}\right) \tag{2-21}$$

式中，$f(x)$ 为普适标度函数，近似表达式为[22]：

$$f(x)\approx\frac{1}{2\sqrt{x^2+2x}}\operatorname{arctanh}\left(\sqrt{\frac{x}{x+2}}\right) \tag{2-22}$$

3．聚类系数

WS 模型的聚类系数为：

$$C_G=\frac{3(K_{\mathrm{nc}}-2)}{4(K_{\mathrm{nc}}-1)}(1-p)^3 \tag{2-23}$$

NW 模型的聚类系数为[23]：

$$C_G=\frac{3(K_{\mathrm{nc}}-2)}{4(K_{\mathrm{nc}}-1)+4K_{\mathrm{nc}}p(p+2)} \tag{2-24}$$

这两个小世界网络模型的聚类系数都保持在一个较大的数值上。

从以上分析可以看出，小世界网络模型具有较小的平均路径长度和较大的聚类系数，可以真实反映现实网络拓扑特性。该模型说明了复杂网络是介于规则网络和随机网络之间的一类网络。但是，该模型无法展现连接度的幂律分布，对真实网络其他特性的模拟相差较远。

2.4.4 无标度网络模型

规则网络、随机网络和小世界网络都是均匀网络，它们的节点度值分布在平均度 $<k>$ 附近。然而实验表明，大多数现实复杂网络并非均匀网络，它们的度分布服从幂律分布，$P(k)\sim k^{-r}$，γ 为常量，反映了度分布的指数规律。而且随着网络规模的不断扩大，具有增长特性，新增节点倾向于与度值大的节点连接。为了解释幂律分布的产生机理，Barabási 和 Albert 提出了一种无标度网络模型[3]，称为 BA（Barabási-Albert）模型。该模型的构造算法如下：

（1）假设网络最初有 m_0 个节点。每次加入一个新节点，新节点通过 $m(m\leqslant m_0)$ 条新加入的边与网络中已有的 m 个节点相连；

（2）新加入节点与已存在节点 v_i 相连接的概率 p_i 正比于节点 v_i 的度 k_i。

将上述步骤重复 t 步后，该算法将会产生一个由 $N_V=m_0+t$ 个节点和 $N_E=mt$ 条边组成的网络。当 $m_0=5$，$m=3$ 时，BA 模型的演化过程如图 2-5 所示。

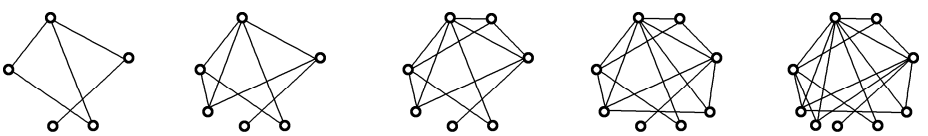


图 2-5　BA 网络模型的演化过程

构造 BA 模型的基本原则包括：

- 增长原则。网络不是一开始就具有大规模特性的，而是通过逐步添加节点和连边形成的；
- 连边倾向原则。新节点倾向于与连接度高的节点相连。BA 模型的拓扑特性主要表现为以下几点。

1．度分布

通过连续场理论方法对 BA 无标度网络的度分布进行计算：假设在 t 时刻，从网络中随机选择一个节点，其连接度为 k 的概率为 $P(k,t)$，称为 t 时刻网络的瞬态度分布。当 $t\to\infty$ 时，如果 $\lim P(k,t)=P(k)$ 存在，则 $P(k)$ 称为网络的稳态度分布。用 $k_i(t)$ 表示在 t 时刻节点 v_i 的度，并假定 $k_i(t)$ 是连续变化的，该方法主要计算 $k_i(t)$ 对 t 的依赖性。每个时间步在增加 m 条边时，节点 v_i 被选中的概率为：

$$m\prod(k_i(t))\left[1-\prod k_i(t)\right]^{m-1}\approx m\prod k_i(t) \tag{2-25}$$

根据连续场理论，$k_i(t)$ 近似满足的动力学方程为：

$$\frac{\partial k_i(t)}{\partial t}=m\prod k_i(t)=m\frac{k_i(t)}{\sum_{j=1}^{N_V}k_j(t)}\approx\frac{k_i(t)}{2t} \tag{2-26}$$

假定 t_i 为节点 v_i 加入网络的时刻，当节点 v_i 在 t_i 时刻加入网络时其度数为 m，即 $k_i(t_i)=m$，因此，动力学方程的解为：

$$k_i(t)=m\left(\frac{t}{t_i}\right)^{\beta} \tag{2-27}$$

式中，$\beta=1/2$ 为动力学指数。由于向网络中增加节点是等时间步长的，因此，t_i 的概率为 $p(t_i)=1/(m_0+t)$，t 时刻网络中任意节点的度值小于 k 的概率为：

$$P(k_i(t)<k)=P\left(t_i>\frac{m^{\frac{1}{\beta}}t}{k^{\frac{1}{\beta}}}\right)=1-P\left(t_i\leqslant\frac{m^{\frac{1}{\beta}}t}{k^{\frac{1}{\beta}}}\right)=1-\frac{m^{\frac{1}{\beta}}t}{k^{\frac{1}{\beta}}(m_0+t)} \tag{2-28}$$

t 时刻网络的瞬态度分布为：

$$P(k,t)=\frac{\partial P(k_i(t)<k)}{\partial k}=\frac{1}{\beta}\frac{m^{\frac{1}{\beta}}t}{(m_0+t)}\frac{1}{k^{\frac{1}{\beta}+1}} \tag{2-29}$$

当 $t\to\infty$ 时，得到网络稳态度分布为：

$$P(k)=\lim_{t\to\infty}P(k,t)\sim 2m^{\frac{1}{\beta}}k^{-r} \tag{2-30}$$

式中，$\gamma=1/\beta+1=3$ 称为度分布指数，是独立于 m 的。由上述公式可知，BA 无标度网络的节点度最终服从指数为 3 的幂律分布。

2．平均路径长度

BA 无标度网络的平均路径长度为[24]：

$$<d>=\frac{\ln N_V}{\ln(\ln N_V)} \tag{2-31}$$

3．聚类系数

BA 无标度网络的聚类系数为[25]：

$$C_G=\frac{m^2(m+1)^2}{4(m-1)}\left[\ln\left(\frac{m+1}{m}\right)-\frac{1}{m+1}\right]\frac{(\ln t)^2}{t} \tag{2-32}$$

以上分析可以看出，BA 模型的平均路径长度较小，聚类系数也较小，但比同规模随

机网络的聚类系数大。经过对各种不同网络模型的研究，BA 模型虽然较好地解释了无标度网络的形成机制，但是它对现实情况的描述过于简化，其演化机制决定了幂律指数近似为一个常数。为了对现实复杂网络进行更深入的研究，还需对 BA 模型进行扩充，考虑更多因素，使它更加符合实际情况。

2.5 传播动力学模型

随着复杂网络研究的快速发展，人们逐步认识到不同事物在真实系统中的传播现象，例如，传染病的流行、计算机病毒的传播以及谣言的扩散等都可以看作是服从某种规律的传播行为。如何描述这些事物的传播过程、揭示它们的传播特性，是传播动力学理论的研究内容。流行病传播数学模型能够很好地描述复杂网络的传播特性，是复杂网络传播动力学的研究基础。

在传播模型的研究中，种群内的个体被抽象为：易感者（Susceptible, S）、感染者（Infected, I）、治愈者（Removed, R）和潜伏者（Exposed, E），个体之间的不同转换构成了不同的传染模型。Kermack 和 McKendrick 构造了著名的 SIR（Susceptible Infected Removed）仓室模型和 SIS（Susceptible Infected Susceptible）仓室模型，在分析所建立模型的基础上，提出了区分流行病的“阈值理论”，为传染病动力学的研究奠定了基础。经典传播模型包括 SI（Susceptible Infected）模型、SIS 模型、SIR 模型以及 SEIR（Susceptible Exposed Infected Removed）模型[26-32]。

1．SI 模型

在 SI 模型中，传播过程只存在易感者和感染者，感染者为传染源，它通过一定概率 v 把传染病传给易感者，易感者一旦被感染，就成为了新的传染源，且被感染个体长期处于感染状态。SI 模型是最简单的传染病传播模型，对于染病后不能治愈的疾病，或者对于突然爆发、尚缺乏有效控制的流行病，在疾病爆发早期常使用 SI 模型进行分析[33]。传播过程如图 2-6 所示。

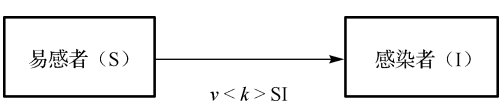


图 2-6　SI 模型传播过程

其中，$<k>$表示人群中的平均连接度，令 $S(t)$ 和 $I(t)$ 分别表示在 t 时刻系统中易感者和感染者数量，N_B 表示 t 时刻人员总数。对 $S(t)$ 和 $I(t)$ 进行归一化处理，令 $S_R(t)$ 和 $I_R(t)$ 分别

表示在 t 时刻易感者和感染者所占人员总数的比例，即 $S_R(t)=S(t)/N_B$，$I_R(t)=I(t)/N_B$，且有 $S_R(t)+I_R(t)=1$。

当时间由 t 变化为 $t+\Delta t$ 时，增加的感染者人数表示为 $I(t+\Delta t)-I(t)$。由于新增加感染者是因为易感者与感染者接触而被传染的，令 λ 为有效传染率，表示单位时间内，一个感染者可以传染的易感者数量，则有：

$$N_B(I_R(t+\Delta t)-I_R(t))=v<k>N_BS_R(t)I_R(t)\Delta t \tag{2-33}$$

公式（2-33）两端同时除以 $N_B\Delta t$，并对 t 求导，可得 SI 模型的动力学方程组为：

$$\begin{cases}\dfrac{\mathrm{d}S_R}{\mathrm{d}t}=-v<k>S_RI_R=-v<k>I_R(1-I_R)\\[2ex]\dfrac{\mathrm{d}I_R}{\mathrm{d}t}=v<k>S_RI_R=v<k>I_R(1-I_R)\\[2ex]I_R(0)=I_0\end{cases} \tag{2-34}$$

式中，I_0 表示在开始时刻感染者所占人员总数的比例。v 为传播概率，反映了疾病本身的传播能力，v 越小，传播速度的最大值越晚到来，v 越大，传播速度的最大值越早到来。由于 SI 模型没有考虑感染者被治愈的情况，传播有效率 $\lambda=v$，当 $t\to\infty$ 时，$I_R(t)\to 1$，表示所有易感者最终都会被感染，成为感染者。

2．SIS 模型

在 SIS 模型中，传播过程只存在易感者和感染者，感染者通过概率 v 把传染病传给易感者。与 SI 模型不同的是，SIS 模型中的感染者本身有一定概率 δ 可以被治愈，$1/\delta$ 表示平均感染期，有效传染率为 $\lambda=v/\delta$。SIS 模型很好地描述了感染个体能够反复被治愈的流行病传播行为[34]。传播过程如图 2-7 所示。

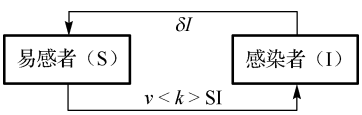


图 2-7　SIS 模型传播过程

假设治愈概率 δ 为固定值，则 SIS 模型的动力学方程组为：

$$\begin{cases}\dfrac{\mathrm{d}S_R}{\mathrm{d}t}=-v<k>S_RI_R+\delta I_R=v<k>(1-S_R)\left(\dfrac{1}{\lambda}-S_R\right)\\[2ex]\dfrac{\mathrm{d}I_R}{\mathrm{d}t}=v<k>S_RI_R-\delta I_R=v<k>I_R(1-I_R)-\delta I_R\end{cases} \tag{2-35}$$

当 $\lambda\leqslant 1/<k>$ 时，方程 $\mathrm{d}S_R/\mathrm{d}t$ 有唯一的平衡点 $S_R(t)=1$，且是渐进稳定的，表示从任意

初值开始，$S_R(t)$ 都将单调增加且趋向于 1，$I_R(t)$ 都将单调减少且趋向于 0，此时疾病不会扩散。而当 $\lambda > 1/<k>$ 时，方程 $\mathrm{d}S_R/\mathrm{d}t$ 有 2 个平衡点：$S_R(t)=1$ 和 $S_R(t)=1/(\lambda<k>)$。$S_R(t)=1$ 时系统不稳定，$S_R(t)=1/(\lambda<k>)$ 时系统渐进稳定。当 $S_R(t)\to 1/(\lambda<k>)$，$I_R(t)\to 1-1/(\lambda<k>)$ 时，感染者始终保持在 N_B（$1-1/(\lambda<k>)$），成为一种地方病。因此，在 SIS 模型中，$\lambda = 1/<k>$ 是区分疾病是否流行的临界值。

3．SIR 模型

在 SIR 模型中，令 $R(t)$ 表示在 t 时刻系统中治愈者数量，$R_R(t)$ 表示在 t 时刻治愈者所占人员总数的比例，则有 $S_R(t)+I_R(t)+R_R(t)=1$。在该模型中，感染者以概率 v 把传染病传给易感者，易感者被感染后成为新的传染源；感染者以概率 δ 被治愈，治愈者对疾病具有免疫能力[35]。传播过程如图 2-8 所示。

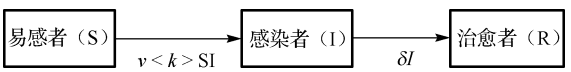


图 2-8　SIR 模型传播过程

SIR 模型的动力学方程组为：

$$\begin{cases} \dfrac{\mathrm{d}S_R}{\mathrm{d}t} = -v<k>S_R I_R \\ \dfrac{\mathrm{d}I_R}{\mathrm{d}t} = v<k>S_R I_R - \delta I_R \\ \dfrac{\mathrm{d}I_R}{\mathrm{d}t} = \delta I_R \end{cases} \tag{2-36}$$

由于 $\mathrm{d}S_R/\mathrm{d}t<0$，$S_R(t)$ 单调递减且有下界。当 $S_R(t)=\delta/(v<k>)$ 时，$I_R(t)$ 达到最大值。当初始时刻的易感者数量 $S_R(0)>\delta/(v<k>)$ 时，感染者数量 $I_R(t)$ 先单调增加达到最大值，然后再逐渐减少。

4．SEIR 模型

在 SEIR 模型中，令 $E(t)$ 表示在 t 时刻系统中潜伏者数量，$E_R(t)$ 表示在 t 时刻潜伏者所占人员总数的比例，则有 $S_R(t)+E_R(t)+I_R(t)+R_R(t)=1$。当易感者被感染后变为潜伏者，潜伏者以概率 p_I 变为感染者，潜伏者不可对易感者进行感染，感染者经过治愈后变为治愈者[36]。传播过程如图 2-9 所示。

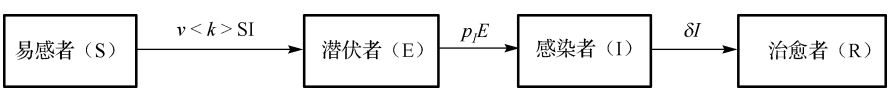


图 2-9　SEIR 模型传播过程

SEIR 模型的动力学方程组为：

$$\begin{cases} \dfrac{\mathrm{d}S_R}{\mathrm{d}t} = -v < k > S_R I_R \\ \dfrac{\mathrm{d}E_R}{\mathrm{d}t} = v < k > S_R I_R - p_I E_R \\ \dfrac{\mathrm{d}I_R}{\mathrm{d}t} = p_I E_R - \delta I_R \\ \dfrac{\mathrm{d}I_R}{\mathrm{d}t} = \delta I_R \end{cases} \tag{2-37}$$

在传播动力学模型中，存在着一个重要阈值，即基本再生数（R_0），它是流行病学中的重要概念，用于评估传染病在人群的最终感染规模[37, 38]。在流行病学中，基本再生数表示一个感染者在平均患病期内所感染的易感者数量[39, 40]，当 $R_0 > 1$ 时，即一个感染者在平均患病期能感染的易感者数量大于 1，那么疾病将始终存在而形成地方病；当 $R_0 < 1$ 时，即一个感染者在平均患病期能感染的易感者数量小于 1，疾病在人群中扩展到一定程度就会自行消亡；$R_0 = 1$ 是判断疾病是否流行的阈值。要防止疾病流行，必须减少 R_0，使它小于 1。

参考文献

[1] S. Boccaletti, V. Latora, Y. Moreno, et al. Complex networks: structure and dynamics[J]. Physics Reports, 2006, 424（4）: 175-308.

[2] Duncan J. Watts, Steven H. Strogatz. Collective dynamics of “small-world” networks[J]. Nature, 1998, 393（6684）: 440-442.

[3] Albert-László Barabási, Réka Albert. Emergence of scaling in random networks[J]. Science, 1999, 286（5439）: 509-512.

[4] Matthew A. Oberhardt, Bernhard O. Palsson, Jason A. Papin. Applications of genome-scale metabolic reconstructions[J]. Molecular Systems Biology, 2009, 5: 320.

[5] Zhi Wang, Jianzhi Zhang. In search of the biological significance of modular structures in protein networks[J]. PLoS Computational Biology, 2007, 3（6）: 1011-1021.

[6] Olaf Sporns, Christopher J. Honey. Small worlds inside big brains[J]. Proceedings of the National Academy of Sciences, 2006, 103（51）: 19219-19220.

[7] 赫南, 淦文燕, 李德毅等. 一个小型演员合作网的拓扑性质分析[J]. 复杂系统与复杂性科学, 2006, 3（4）: 1-10.

[8] M. E. J. Newman. Scientific coauthorship networks and patterns of scientific collaboration[J]. Proceedings of the National Academy of Sciences, 2004, 101（51）: 5200-5205.

[9] James Moody. The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999[J]. American Sociological Review, 2004, 69（2）: 213-238.

[10] Holger Ebel, Lutz I. Mielsch, Stefan Bornholdt. Scale-free topology of e-mail networks[J]. Physical Review E, 2002, 66: 035103.

[11] Ryan Rowe, Germ´an Creamer. Automated social hierarchy detection through email network analysis[C]. Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis, 2007: 109-117.

[12] Yanbin Mao, Feng Liu, Shengwei Mei. On the topological characteristics of power grids with distributed generation[C]. Proceedings of the 29th Chinese Control Conference, 2010: 4714-4720.

[13] 杨洪勇, 路兰, 张嗣瀛. 基于复杂网络的 Internet 结构模型[J]. 控制工程, 2010, 17（3）: 380-383.

[14] Albert-László Barabási, Réka Albert, Hawoong Jeong. Scale-free characteristics of random networks: The topology of the World Wide Web[J]. Physica A: Statistical Mechanics and its Applications, 2000, 281（1）: 69-77.

[15] Fang Wang, Yamir Moreno, Yaoru Sun. Structure of Peer-to-Peer social networks[J]. Physical Review E, 2006, 73（3）: 036123.

[16] Marián Boguñá, Romualdo P. Satorras, Alessandro Vespignani. Epidemic spreading in complex networks with degree correlations[J]. Lecture Notes in Physics, 2003, 625: 127-147.

[17] Tao Zhou, Zhongqian Fu, Binghong Wang. Epidemic dynamics on complex networks[J]. Progress in Natural Science, 2006, 16（5）: 452-457.

[18] M. E. J. Newman. Assortative mixing in networks[J]. Physical Review Letters, 2002, 89（20）: 208701.

[19] M. E. J. Newman, Duncan J. Watts. Scaling and percolation in the small-world network model[J]. Physical Review E, 1999, 60（6）: 7332-7342.

[20] A. Barrat, M. Weigt. On the properties of small-world models[J]. The European Physical Journal B - Condensed Matter and Complex Systems, 2000, 13（3）: 547-560.

[21] M. E. J. Newman, Duncan J. Watts. Renormalization group analysis of the small-world network model[J]. Physics Letters A, 1999, 263（4）: 341-346.

[22] M. E. J. Newman, Cristopher Moore, Duncan J. Watts. Mean field solution of the small-world network model[J]. Physical Review Letters, 2000, 84（14）: 3201-3204.

[23] M. E. J. Newman. The structure and function of networks[J]. Computer Physics Communications, 2002, 147（1）: 40-45.

[24] Reuven Cohen, Shlomo Havlin. Scale-free networks are ultrasmall[J]. Physical Review Letters, 2003, 90

（5）: 058701.

[25] Agata Fronczak, Piotr Fronczak, Janusz A. Holyst. Mean-field theory for clustering coefficients in Barabási-Albert networks[J]. Physical Review E, 2003, 68（4）: 046126.

[26] 马知恩, 周义仓, 王稳地等. 传染病动力学的数学建模与研究[M]. 北京: 科学出版社, 2004.

[27] 杨伟. 传染病动力学的一些数学模型及其分析[D]. 复旦大学博士学位论文, 2010.

[28] Gang Yan, Tao Zhou, Jie Wang, et al. Epidemic spread in weighted scale-free networks[J]. Chinese Physics Letters, 2005, 22（2）: 510.

[29] Matt J. Keeling, Ken T. D. Eames. Networks and epidemic models[J]. Journal of the Royal Society Interface, 2005, 2（4）: 295-307.

[30] A. Ganesh, L. Massoulie, D. Towsley. The effect of network topology on the spread of epidemics[C]. Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies, 2005, 2: 1455-1466.

[31] Xiaoyan Wu, Zonghua Liu. How community structure influences epidemic spread in social networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387（2）: 623-630.

[32] Marián Boguñá, Romualdo P. Satorras. Epidemic spreading in correlated complex networks[J]. Physical Review E, 2002, 66（4）: 047104.

[33] Marc Barthélemy, Alain Barrat, Romualdo P. Satorras, et al. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks[J]. Physical Review Letters, 2004, 92（17）: 178701.

[34] Chengyi Xia, Zhongxin Liu, Zengqiang Chen, et al. Spreading behavior of SIS model with non-uniform transmission on scale-free networks[J]. The Journal of China Universities of Posts and Telecommunications, 2009, 16（1）: 27-31.

[35] J. Satsuma, R. Willox, A. Ramani, et al. Extending the SIR epidemic model[J]. Physica A: Statistical Mechanics and its Applications, 2004, 336（3）: 369-375.

[36] Gergely Rost, Jianhong Wu. SEIR epidemiological model with varying infectivity and infinite delay[J]. Mathematical Biosciences and Engineering, 2008, 5（2）: 389-402.

[37] Junling Ma, David J. D. Earn. Generality of the final size formula for an epidemic of a newly invading infectious disease[J]. Bulletin of Mathematical Biology, 2006, 68（3）: 679-702.

[38] Claudio Castellano, Romualdo P. Satorras. Thresholds for epidemic spreading in networks[J]. Physical Review Letters, 2010, 105（21）: 218701.

[39] J. M. Heffernan, R. J. Smith, L. M. Wahl. Perspectives on the basic reproductive ratio[J]. Journal of the Royal Society Interface, 2005, 2（4）: 281-293.

[40] Matt Keeling. The implications of network structure for epidemic dynamics[J]. Theoretical Population Biology, 2005, 67（1）: 1-8.

第3章

P2P 网络特定信息传播模型

3.1 引言

在 1.2 节中，对 P2P 网络及其信息传播模式进行了介绍。从中可以看出，P2P 文件共享系统在方便人们共享文件的同时，也带来了严重的信息安全问题，在 P2P 网络上充斥着大量的色情、暴力、迷信、反华宣传、涉密信息、盗版文件等不良信息，给社会和谐稳定、网络文化安全、知识产权保护以及国家安全带来极大的危害。

由于 P2P 网络采用对等计算模式和非标准通信协议，对 P2P 网络信息安全监控提出很大的挑战。国内外对 P2P 网络信息安全监控技术开展了大量的研究工作，提出了多种 P2P 网络信息监控方法，归纳起来可分成被动监控方法和主动控制方法两种。

被动监控方法主要采用基于 P2P 流量特征识别的整体流量封堵技术，该方法并不适合对 P2P 特定信息传播的监控，其原因如下：

（1）整体流量封堵技术通常作为网络营运商的 P2P 网络流量管理手段，通过封堵或限制某种 P2P（如 BitTorrent、eMule 等）网络流量，实现对网络带宽的管理，在封堵时并不区分所传输的内容。并且目前的 P2P 系统大多数采用动态端口分配和数据加密技术，使得基于 P2P 流量特征识别的 P2P 网络监控技术面临很大的挑战。

（2）对于互联网中的海量网络流量数据，在实施网络流量检测和识别时会引入较大的网络延迟，造成网络服务质量的下降。

（3）P2P 系统通常工作在网络边缘，通过边缘节点之间的协作实现文件资源共享。而网络营运商的 P2P 网络流量监控设施通常部署在网络核心，难以监控到网络边缘节点之间的 P2P 文件共享行为。

（4）对于网络安全监管部门来说，主要关注的是 P2P 网络特定内容信息传播及其监控问题，与网络营运商所关注的角度不同，要求能够对 P2P 网络中特定文件下载或特定信息传播行为进行有效监控，而被动监控技术难以满足这一需求。

主动控制方法主要采用基于文件污染的 P2P 特定文件下载控制技术，该方法首先被

美国唱片工业所采用，用于对网络音像作品版权保护中。该方法的基本原理是在传播盗版文件的 P2P 网络中，对被盗版的文件实施污染，即在被盗版的共享文件中有意加入错误的内容，将这些被污染的文件作为污染源放入 P2P 网络中诱使用户下载。由于用户不能分辨出文件是否被污染，便将这些被污染的文件下载到他们的主机中，并与其他用户共享，于是这些被污染的文件便在 P2P 网络中传播开。由于被污染的文件对用户的文件下载过程产生干扰，使之难以完成文件下载任务，白白浪费用户的宝贵时间和网络带宽，造成用户对该 P2P 网络环境的不信任感，进而离开该 P2P 网络不再去下载盗版文件，遏制了盗版文件在该 P2P 网络中的传播，促使用户购买和使用正版文件。例如，美国一家专业的 P2P 污染公司[1]曾在 FastTrack（一种 P2P 文件共享网络）上发布共享文件的污染版本，占据了该网络中所有共享文件的一半以上，在一定程度上遏制了盗版文件的传播。然而，由于现在的 P2P 软件大都引入了信誉机制或修复机制等反污染技术，以对抗文件污染，削弱了文件污染的控制效果。

随着 P2P 文件共享技术的广泛应用，网络信息安全问题变得日益突出。由于缺乏有效的监控手段，目前还难以对 P2P 网络中传播的不良信息进行有效控制，给网络信息安全和文化安全带来潜在的隐患和风险。根据 P2P 网络信息传播的自身特点和规律性，主动控制方法比较适合对 P2P 特定信息传播的控制，也是国内外重点研究和发展的 P2P 信息传播控制技术。

本章给出一种主动控制方案，通过对 P2P 信息传播过程的建模分析，研究 P2P 信息传播特性和内在规律，寻找影响 P2P 特定信息传播的关键节点，通过控制这些关键节点，实现对 P2P 特定信息传播的主动控制。

3.2　P2P 网络测量模型

P2P 网络测量是 P2P 网络信息传播特性分析的基础，通过网络测量，采集大量的 P2P 网络数据，为分析和验证 P2P 网络信息传播特性提供数据资源。按照测量方法分类，P2P 网络测量模型可分成主动测量模型和被动测量模型两种。

3.2.1　主动测量模型

主动测量方法是将用于网络测量和数据采集的测量节点加入到 P2P 网络中，主动地测量和采集相关的 P2P 网络数据，如 P2P 节点的 IP 地址、端口号以及所有可以通过 P2P 通信协议获取的“元信息”等，这些信息可用于分析和监测 P2P 网络的拓扑、延迟、内容可用性、节点特性等参数。

测量节点通常采用仿真 P2P 客户端技术来实现。通常，一个用户主机要加入到一个

特定的 P2P 网络（如 BitTorrent 等）中，首先需要运行该网络的客户端软件，成为该网络的一个 P2P 节点，然后才能获得该 P2P 网络所提供的服务或资源。所谓仿真客户端，是指按照特定的 P2P 客户端软件的系统架构和通信协议开发的测量客户端，它能够像普通 P2P 节点一样加入到 P2P 网络中，主要用于测量和采集相关的 P2P 网络数据。在一般情况下，测量客户端是通过改造开源 P2P 客户端实现的，主要用于支持网络测量和数据采集任务。

下面给出一个主动测量模型及实现方案。

1．模型框架

该模型主要用于测量和采集在 P2P 网络中参与传播特定文件的受众信息，以 BitTorrent 网络为监测对象。该模型主要由“特定信息”主题管理、“元信息”搜索、节点列表获取、节点状态及连接关系获取等部分组成，模型框架如图 3-1 所示。

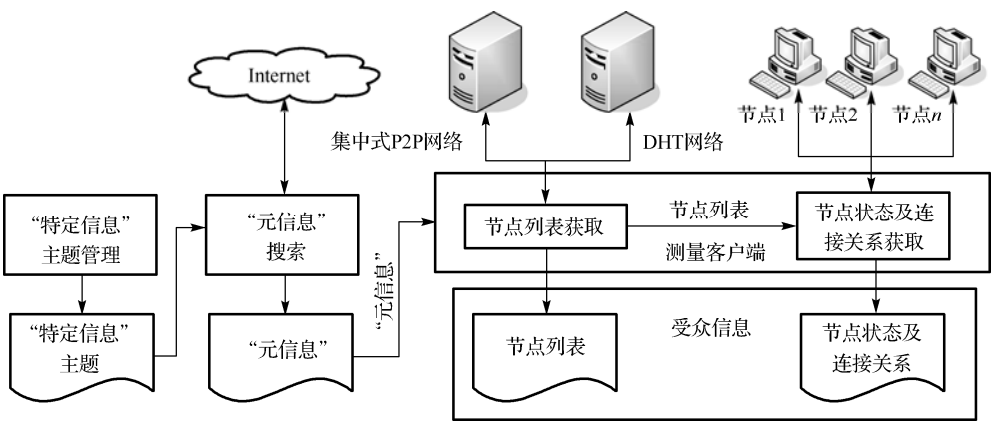


图 3-1　主动测量模型框架

模型中的主要模块功能描述如下：

（1）“特定信息”主题管理。主要对需要关注的“特定信息”主题进行管理，包括敏感主题、非法主题、侵权主题等，为“元信息”搜集做准备。

（2）“元信息”搜索。“元信息”是指启动一个 P2P 特定文件传播任务所需的基本信息，其中包括文件列表或文件名、文件大小、文件 Hash 值、服务器地址列表、端口号等。BitTorrent 中的 Torrent 种子文件就是典型的“元信息”。“元信息”搜索主要通过主题网络爬虫工具从互联网上获取与“特定信息”主题相关的所有“元信息”。获取的“元信息”是测量客户端进行受众信息采集的基础。

（3）节点列表获取。通过节点列表来获取受众信息，包括文件在 P2P 网络中传播时所涉及的节点信息、节点状态信息以及节点之间的连接信息。节点列表获取模块就是在测

量客户端中通过模拟 BitTorrent 所使用的通信协议，向 BitTorrent 网络中的相关节点（如 Tracker 服务器和 DHT 入口节点）发送仿真节点列表获取数据包，并对返回数据包进行分析，获取节点列表信息。

（4）节点状态及连接关系获取。当测量客户端得到节点列表后，为了进一步了解节点状态信息，需要通过模拟节点之间的交互过程，得到相应节点的状态信息和对特定信息的资源拥有情况。在与节点交互的同时，可以对测量客户端与节点之间的距离进行测量；同时通过使用 PEX（Peer Exchange）技术获取对方节点正在连接的邻居节点列表信息，建立起节点之间的连接关系和网络拓扑结构。

该模型以 P2P 特定信息传播及其受众为监测目标，对相关数据进行测量和采集，并且引入了 PEX 技术，不仅可以获得节点之间的连接关系，还提高了受众信息的获取效率。

2．PEX 技术

对大规模 P2P 网络进行主动测量时，由于很难大范围地部署测量节点，因此难以得到其他 P2P 节点之间的连接信息和网络拓扑信息。针对这个问题，可以通过 P2P 协议中的 PEX 技术来获取节点之间的连接关系，并构建基于特定信息的 P2P 传播网络拓扑结构。

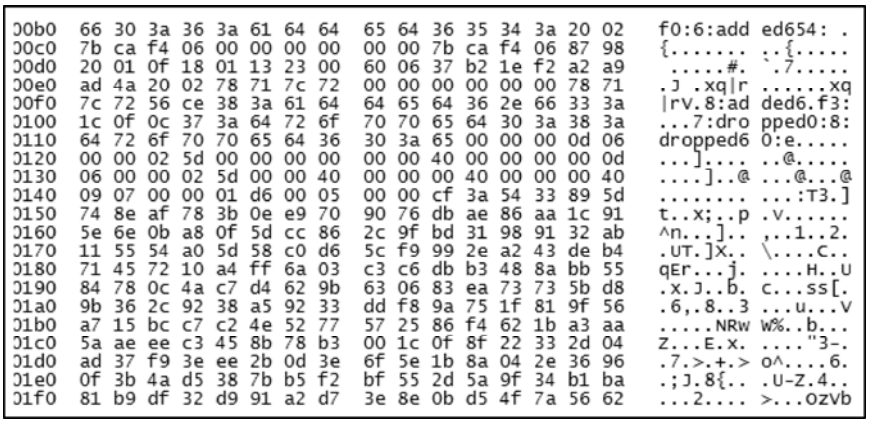
在最初的 BitTorrent 协议中，用于文件传输的节点列表是通过 Tracker 服务器或 DHT 网络获得的，由于 Tracker 服务器的集中性，容易被封堵。后来的 BitTorrent 协议采用了 PEX 技术，使节点直接与其他节点交换各自拥有的邻居节点信息，减少对 Tracker 服务器和 DHT 网络的依赖，使 BitTorrent 网络更加高效、迅速和鲁棒。

PEX 消息不能独立地工作，请求消息作为握手协议的扩展部分进行发送，响应消息只有当节点间采用 Peer Wire 协议建立连接时，才会定期在连接中传输。因此，初始连接节点必须使用传统方式获取。PEX 包括两种扩展协议：一是基于 AZMP 的 AZ_PEX；二是基于 LTEP 的 UT_PEX。这两种扩展协议的返回消息都采用 Bencode 编码，包含一组增加的节点列表和一组删除的节点列表。关键字 added 后紧跟的是在连续 2 次 PEX 消息之间新增连接的节点列表，关键字 dropped 后紧跟的是在连续 2 次 PEX 消息之间断开连接的节点列表。节点列表组成格式为每 6 个字节表示一个节点信息，前 4 个字节表示 IP 地址，后 2 个字节表示端口号，如图 3-2 所示。

在客户端中，PEX 消息必须符合如下两项要求：

（1）为了减少消息长度，每个消息中增加的节点数量和删除的节点数量不超过 50 个，如果超过 50 个，则要分成多个消息进行发送；

（2）为了防止节点之间的 PEX 消息形成泛洪式发送，节点之间发送 PEX 消息的间隔时间需要超过 1 分钟。



```
00b0  66 30 3a 36 3a 61 64 64  65 64 36 35 34 3a 20 02   f0:6:add ed654: .
00c0  7b ca f4 06 00 00 00 00  00 00 7b ca f4 06 87 98   {....... ..{.....
00d0  20 01 0f 18 01 13 23 00  60 06 37 b2 1e f2 a2 a9    .....#. `.7.....
00e0  ad 4a 20 02 78 71 7c 72  00 00 00 00 00 00 78 71   .J .xq|r ......xq
00f0  7c 72 56 ce 38 3a 61 64  64 65 64 36 2e 66 33 3a   |rV.8:add ded6.f3:
0100  1c 0f 0c 37 3a 64 72 6f  70 70 65 64 30 3a 38 3a   ...7:dro pped0:8:
0110  64 72 6f 70 70 65 64 36  30 3a 65 00 00 00 0d 06   dropped6 0:e.....
0120  00 00 02 5d 00 00 00 00  00 00 40 00 00 00 00 0d   ...]....  ..@.....
0130  06 00 00 02 5d 00 00 40  00 00 00 40 00 00 00 40   ....]..@ ...@...@
0140  09 07 00 00 01 d6 00 05  00 00 cf 3a 54 33 89 5d   ........ ...:T3.]
0150  74 8e af 78 3b 0e e9 70  90 76 db ae 86 aa 1c 91   t..x;..p .v......
0160  5e 6e 0b a8 0f 5d cc 86  2c 9f bd 31 98 91 32 ab   ^n...]..  ,..1..2.
0170  11 55 54 a0 5d 58 c0 d6  5c f9 99 2e a2 43 de b4   .UT.]X..  \....C..
0180  71 45 72 10 a4 ff 6a 03  c3 c6 db b3 48 8a bb 55   qEr...j.  ....H..U
0190  84 78 0c 4a c7 d4 62 9b  63 06 83 ea 73 73 5b d8   .x.J..b.  c...ss[.
01a0  9b 36 2c 92 38 a5 92 33  dd f8 9a 75 1f 81 9f 56   .6,.8..3  ...u...V
01b0  a7 15 bc c7 c2 4e 52 77  57 25 86 f4 62 1b a3 aa   .....NRw  W%..b...
01c0  5a ae ee c3 45 8b 78 b3  00 1c 0f 8f 22 33 2d 04   Z...E.x.  ....".3-.
01d0  ad 37 f9 3e ee 2b 0d 3e  6f 5e 1b 8a 04 2e 36 96   .7.>.+.>  o^....6.
01e0  0f 3b 4a d5 38 7b b5 f2  bf 55 2d 5a 9f 34 b1 ba   .;J.8{..  .U-Z.4..
01f0  81 b9 df 32 d9 91 a2 d7  3e 8e 0b d5 4f 7a 56 62   ...2....  >...OzVb
```

图 3-2　PEX 返回消息内容

虽然 PEX 技术还不能完全取代传统节点列表获取方法，还必须使用传统 P2P 技术获取开始连接节点，但是随着 PEX 技术的应用，能够大幅度减少对 Tracker 服务器和 DHT 网络的依赖，提高节点获取效率和 P2P 网络的鲁棒性。

3．受众信息获取技术

该模型通过测量客户端（即测量节点）对 P2P 文件共享系统中的特定信息传播过程进行监测，以获得 P2P 特定信息传播过程中的受众信息，如节点信息、节点状态信息以及节点之间的连接关系等，构造 P2P 特定信息传播在某一时刻的静态网络拓扑和在一段时间内的动态变化网络拓扑，分析 P2P 特定信息的传播规律。

1）节点列表获取

节点列表获取是根据被监测 P2P 特定信息的“元信息”内容，模拟 P2P 协议中的节点列表请求消息，发送给 Tracker 服务器和 DHT 网络入口节点，并解析响应消息，得到节点列表。下面以 BitTorrent 协议和 DHT 协议为例，介绍测量客户端与 Tracker 服务器和 DHT 网络的交互过程。

BitTorrent 协议主要用于节点与 Tracker 服务器进行信息交互，当节点需要下载特定文件时，向 Tracker 服务器发送节点列表请求消息，Tracker 服务器接收到请求消息后，根据文件 Hash 查询符合要求的节点列表，并随机选择部分节点生成响应消息发送给请求节点，请求节点得到响应消息后，对消息内容进行解析，得到节点列表。节点与 Tracker 服务器之间的消息交互通过 HTTP 协议实现，请求为 Get 方式，响应为 Response 方式。

请求节点列表的消息格式为：TrackerAddress?info_hash= &peer_id= &ip= &port= &uploaded= &downloaded= &left= &numwant= &compact=1&event=started，具体参数含义如下。

（1）TrackerAddress 是 Tracker 服务器地址，一般以 http://开始；

（2）info_hash 表示文件 Hash，长度为 20 字节，生成消息时需要转义；

（3）peer_id 表示当前节点 ID，长度为 20 字节，生成消息时需要转义；

（4）ip 和 port 表示当前节点的 IP 地址和端口号；

（5）uploaded、downloaded 和 left 分别表示已上传、已下载和未下载的数据大小；

（6）numwant 和 compact 分别表示期望返回节点数量和返回消息是否压缩；

（7）event 的选项有 started，completed 和 stopped，表示节点请求状态。

Tracker 服务器响应的 Response 消息格式为：peers:***。其中，peers:为关键字，*** 表示节点列表信息，节点列表的组成格式为每 6 个字节表示一个节点信息，前 4 个字节表示节点 IP 地址，后 2 个字节表示节点端口号。

DHT 协议主要用于在 DHT 网络中获取节点列表，请求消息和响应消息采用 UDP 协议来传送。在请求节点列表时，首先向每个 DHT 入口节点发送请求消息，请求消息格式为：d1:ad2:id20: 9:info_hash20:info_hashe1:q9:get_peers1:t8:******** 1:y1:qe。其中，info_hash 为“元信息”中的文件 Hash；********表示 8 位随机字符串；其他内容为固定格式。

接收到请求消息的节点根据 DHT 协议查询节点列表，并返回查询结果，响应消息格式为：d1:rd2:id20%%%nodes:***values:***。其中，d1:rd2:id20 为固定格式，用于判断响应消息是否可用；%%%为辅助信息；nodes 和 values 是关键字，nodes 表示后续节点不是目标节点，测量客户端还需进一步请求，将 nodes 后面的节点信息放入待请求节点列表中，values 表示后续节点为目标节点，可以直接与其连接进行文件传输；***表示节点列表信息。响应消息解析完成后，测量客户端需要从待请求节点列表中取出部分节点，向其发送请求消息并解析响应消息，直到待请求节点列表为空为止。

2）节点状态信息获取

在获取节点列表后，需要进一步了解节点状态以及节点间的距离。节点状态信息包括：节点是否在线、节点 ID 信息、节点对特定信息的资源拥有情况等。下面以 Peer Wire 协议为例，介绍测量客户端与节点的交互过程。

节点状态获取主要通过 Peer Wire 协议来实现，Peer Wire 协议是基于 TCP 的应用层协议。节点之间的连接通过 Peer Wire 协议中的握手消息开始，握手消息格式为：0X13 BitTorrent protocol 00000000 info_hash peer_id。其中，0X13 为消息的第一个字节内容，表示关键字 BitTorrent protocol 的长度；BitTorrent protocol 为 Peer Wire 协议关键字；00000000 为保留字节；info_hash 表示“元信息”中的文件 Hash；peer_id 表示当前节点 ID。当对方节点收到握手请求消息后，如果该节点拥有相关文件信息，则会及时发送响应消息，否则将丢弃握手请求消息，不做响应。响应消息的格式为：0X13 BitTorrent

protocol 00000000 info_hash peer_id ####。其中，peer_id 表示对方节点 ID；####表示对方节点对特定信息的资源拥有情况，大小为 4 个字节，将这 4 个字节的内容转换为二进制，二进制中的每一位表示对相应文件块（Piece）的拥有情况，1 表示拥有，0 表示不拥有。

P2P 节点间距离测量比较复杂，因为 P2P 网络是一种叠加在 IP 网络上的逻辑覆盖网络，节点距离很近的节点在实际物理网络中有可能相距很远，而距离很远的节点在实际物理网络中反而有可能相距很近。为了测量 P2P 节点之间的实际物理距离，测量客户端采用如下的测量方法：首先采用 ICMP 协议中的 Ping 操作进行测量，使用应答包返回的生存时间（Time To Live，TTL）和环路时延（Round-Trip Time，RTT）来表示距离，研究表明，RTT 和 TTL 可以用来描述实际 P2P 网络中两个节点之间的距离[2]。由于在网络中大量使用了网络地址翻译（Network Address Translator，NAT）技术和动态主机配置协议（Dynamic Host Configuration Protocol，DHCP），破坏了 IP 地址与 P2P 节点之间的一一对应关系，因此测量结果不但要与节点 IP 地址相对应，还需要与节点 peer_id 联系起来，将三者结合起来，能够准确地获得 P2P 节点的 IP 地址变化情况。

3）节点连接关系获取

测量客户端主要采用 PEX 技术来获取 P2P 特定信息传播网络中节点之间的连接关系。根据 PEX 技术特点，当测量客户端与节点进行握手操作时，由测量客户端将 PEX 消息内容作为扩展消息附加到握手消息后，发送给对方节点。如果对方节点也支持 PEX 技术，则会随机选取正在连接的邻居节点进行响应，否则不响应 PEX 消息。如果测量客户端在一段时间内没有得到对方节点的 PEX 响应消息，则认为该节点不支持 PEX 消息，将该节点列入“PEX 黑名单”中，以后不再向该节点发送 PEX 扩展消息。

测量客户端对 PEX 响应消息进行解析，得到需要增加的节点列表和需要删除的节点列表，从而获取对方节点与其他节点之间的连接关系。由于测量客户端得到的节点数量庞大，如果与每个节点都建立 P2P 连接的话，系统资源会很快被消耗殆尽。因此，需要及时断开已经获得响应消息的 P2P 连接以及长时间得不到响应消息的 P2P 连接。由此带来的问题是每次得到的节点列表都是对方节点正在连接的那部分节点，因此需要对一段时间内的新增节点和删除节点进行统计。

PEX 通常包括两种扩展协议：AZ_PEX 和 UT_PEX，常用的是 UT_PEX。在 UT_PEX 扩展协议中，PEX 请求消息格式为：d1:md11:LT_metadatai1e6:µt_pex i2ee1:pi6881e1:v13:\xc2\xb5Torrent 1.2e。其中，d11:LT_metadatai1 表示握手消息支持扩展协议；µt_pex 为 PEX 扩展消息关键字；pi6881 表示当前节点的端口号为 6881；v13:\xc2\xb5Torrent 1.2e 表示当前客户端的版本为“µTorrent 1.2”。当握手消息中包含 PEX 请求消息时，保留字段中的第 6 位内容为 10，第 8 位内容为 05。

4．模型性能分析

下面通过实验方法来分析在模型中引入 PEX 技术后对受众信息获取效率和节点覆盖率的影响。一个设计良好的 P2P 主动测量模型应当具有较高的测量效率和节点覆盖率，并尽量减少向 P2P 网络的注入流量。在主动测量模型中，如果不考虑 PEX 技术的使用，对特定文件的测量步骤主要包括：节点列表获取、节点状态获取。引入 PEX 技术后，对测量过程有如下几点影响：

（1）向已获取节点发送 Peer Wire 握手消息时，需要生成 PEX 扩展消息并附加到握手消息中进行发送。

（2）等待对方节点的 PEX 响应消息，并对响应消息进行解析，得到对方节点的邻居节点列表，建立对方节点与其他节点之间的连接关系。

（3）将通过 PEX 技术得到的邻居节点列表作为受众信息加入到已获取节点列表中。

PEX 技术的使用，不但获得了节点之间的连接关系，而且提高了节点列表的获取能力。假设 t_i 表示测量客户端第 i 次获取受众信息的时间点，$N_{\mathrm{B}i}$ 为该时间点通过 Tracker 服务器获取的节点数量，$N_{\mathrm{D}i}$ 为该时间点通过 DHT 网络获取的节点数量，N_{D} 为 DHT 网络的节点数量，$N_{\mathrm{N}i}$ 表示在时间点 t_i 所有参与特定信息传输的节点数量，在时间充裕的情况下，$N_{\mathrm{D}i} \approx N_{\mathrm{N}i}$。在不使用 PEX 技术时，第 i 次节点列表获取过程所涉及的消息数量 $N_{\mathrm{M}i}$ 为：

$$N_{\mathrm{M}i} = 2 + 2F(N_{\mathrm{B}i}, N_{\mathrm{D}i}) + 2N_{\mathrm{D}i}O(\ln N_{\mathrm{D}}) \tag{3-1}$$

式中，$O(\ln N_{\mathrm{D}})$ 表示在 DHT 网络中返回一个节点所需要的查询次数；$F(N_{\mathrm{B}i}, N_{\mathrm{D}i})$ 表示对 $N_{\mathrm{B}i}$ 个节点和 $N_{\mathrm{D}i}$ 个节点过滤重复节点后的节点数量。当使用 PEX 技术时，第 i 次节点列表获取过程所涉及的消息数量 $N_{\mathrm{ME}i}$ 为：

$$N_{\mathrm{ME}i} = 2 + 2F(N_{\mathrm{B}i}, N_{\mathrm{D}i}) + 2N_{\mathrm{D}i}O(\ln N_{\mathrm{D}}) + p_{\mathrm{pex}}F(N_{\mathrm{B}i}, N_{\mathrm{D}i}) \tag{3-2}$$

式中，p_{pex} 表示节点支持 PEX 技术的概率。可以看出，使用 PEX 技术后，增加的消息数量最多为 $p_{\mathrm{pex}}F(N_{\mathrm{B}i}, N_{\mathrm{D}i})$。

在通过 Tracker 服务器获取节点列表时，节点被选中并返回的概率为：

$$p_{\mathrm{S}i} = \begin{cases} \dfrac{N_{\mathrm{B}i}}{N_{\mathrm{N}i}} & N_{\mathrm{N}i} > N_{\mathrm{B}i} \\ 1 & N_{\mathrm{N}i} \leqslant N_{\mathrm{B}i} \end{cases} \tag{3-3}$$

式中，$p_{\mathrm{S}i}$ 表示节点被返回的概率。当 $N_{\mathrm{N}i} > N_{\mathrm{B}i}$ 时，有部分节点没有被返回，假设共进行了 N_T 次获取，节点始终没有被选中并返回的概率 p_{NS} 为：

$$p_{NS}=\prod_{i=1}^{N_T}\left(1-\frac{N_{\mathrm{B}i}}{N_{\mathrm{N}i}}\right) \tag{3-4}$$

因此，通过 Tracker 服务器方式获取节点列表的整体覆盖率为：

$$g_{fS}=1-p_{NS}=1-\prod_{i=1}^{N_T}\left(1-\frac{N_{\mathrm{B}i}}{N_{\mathrm{N}i}}\right) \tag{3-5}$$

在 BitTorrent 协议中，每次返回的节点数量是有上限的。为了提高覆盖率，可以采用增加获取次数的方法来提高返回节点数量，同时也会增加网络测量的消息发送量。

当通过 DHT 网络获取节点列表时，理论上可以获取到与特定信息相关的所有节点列表，但是每获取一个节点都需要与平均 $O(\ln N_{\mathrm{D}})$ 个节点进行交互，所需时间和消息数量较多。

当使用 PEX 技术时，节点未被搜索到的情况为：与 Tracker 服务器交互时，该节点未被返回；该节点只与不支持 PEX 技术的节点相连。因此，通过 PEX 技术得到的节点覆盖率为：

$$g_{\mathrm{fX}i}=1-\left(1-\frac{N_{\mathrm{B}i}}{N_{\mathrm{N}i}}\right)(1-p_{\mathrm{pex}})^{N_{\mathrm{N}i}(1-g_{\mathrm{fX}i})} \tag{3-6}$$

求解公式（3-6）可以得到 $g_{\mathrm{fX}i}$。可以看出，p_{pex} 是提高节点覆盖率的关键，在现有 BitTorrent 软件中，PEX 技术是默认启用的，因此使用 PEX 技术可以得到较高的节点覆盖率。

使用 PEX 技术完成一次搜索所需要的时间为：

$$T_{\mathrm{X}i}=T_{\mathrm{Server}}+N_{\mathrm{N}i}g_{\mathrm{fX}i}T_{\mathrm{Peer}} \tag{3-7}$$

式中，T_{Server} 表示与 Tracker 服务器交互一次所花费的平均时间，T_{Peer} 表示与节点交互一次所花费的平均时间。

下面通过一个实验对 PEX 技术的使用效果进行分析和验证。在实验中，使用主动测量模型对视频文件“Just Do With It”进行测量，向 Tracker 服务器和 DHT 网络循环获取节点列表的间隔时间设置为 60 秒，图 3-3 显示了通过 Tracker 服务器、DHT 网络和 PEX 技术获取节点数量的变化情况。

从图 3-3 可以看出，Tracker 服务器的节点返回速度较快，但是由于返回的节点列表是 Tracker 服务器从缓存中随机选择的，存在很多重复节点，所以节点数量增长较慢；而 DHT 网络的节点返回速度开始时较慢，后期返回速度明显加快；通过 PEX 技术获取节点列表时，获取速度较快、数量较多。可以看出，通过 PEX 技术的使用，能够大幅度提高节点列表获取的效率和速度。

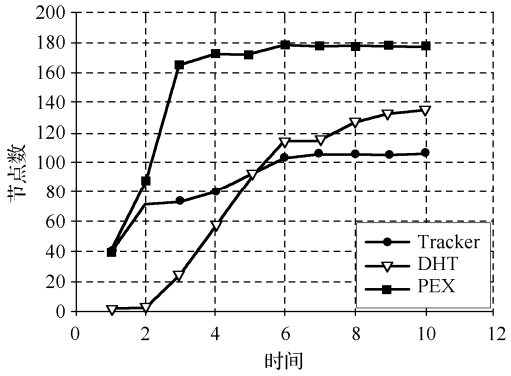


图 3-3　节点列表获取速度对比

在一段时间内，$\Phi_{\mathrm{PB1}},\Phi_{\mathrm{PB2}},\cdots,\Phi_{\mathrm{PB}i},\cdots$为通过 Tracker 服务器获取的节点列表集合，$\Phi_{\mathrm{PD1}},\Phi_{\mathrm{PD2}},\cdots,\Phi_{\mathrm{PD}j},\cdots$为通过 DHT 网络获取的节点列表集合，$\Phi_{\mathrm{PX1}},\Phi_{\mathrm{PX2}},\cdots,\Phi_{\mathrm{PX}k},\cdots$为通过使用 PEX 技术获取的节点列表集合，将获取的所有节点进行合并，形成集合 $\Phi_{\mathrm{P}}=\Phi_{\mathrm{PB1}}\cup\cdots\cup\Phi_{\mathrm{PB}i}\cup\cdots\cup\Phi_{\mathrm{PD1}}\cup\cdots\cup\Phi_{\mathrm{PD}j}\cup\cdots\cup\Phi_{\mathrm{PX1}}\cup\cdots\cup\Phi_{\mathrm{PX}k}\cup\cdots$。由于难以得到实际网络中的所有节点列表信息，可使用集合 Φ_{P} 近似代替实际网络中的所有节点列表，对节点覆盖率进行分析。图 3-4 显示了通过 Tracker 服务器、DHT 网络和 PEX 技术获取节点列表时的节点覆盖率变化情况。

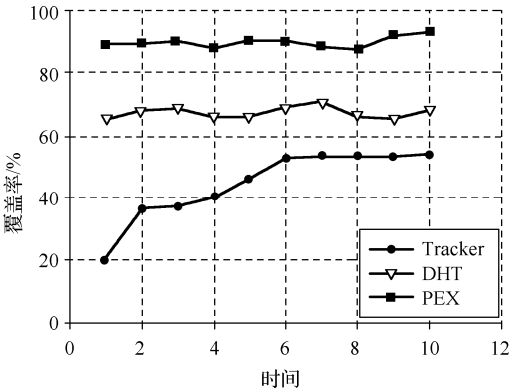


图 3-4　节点覆盖率对比

从图 3-4 可以看出，通过 Tracker 服务器获取节点的覆盖率随着获取次数的增加而缓慢增长，最终趋近于中间水平，主要原因是通过 Tracker 服务器无法得到 DHT 网络中的节点信息；通过 DHT 网络获取节点的覆盖率随着获取次数的增加始终保持在稳定水平，由于 DHT 网络无法返回只通过 Tracker 服务器进行文件交换的节点，节点覆盖率并不

高；通过 PEX 技术获取节点的覆盖率随着获取次数的增加始终保持在较高水平。由此可以看出，通过 PEX 技术的使用，能够将节点覆盖率大幅度提高，并且始终保持在稳定水平。

通过对主动测量模型的性能分析可以看出，在主动测量模型中引入 PEX 技术，不仅可以得到节点之间的连接关系，而且能大幅度提高受众信息的获取速度和覆盖率，从而提高主动测量模型的测量效率。

3.2.2 被动测量模型

被动测量方法通常在网络不同位置部署一定数量的测量点，使用特定的软件或硬件设备采集 P2P 流量数据来实施网络测量。为了保证测量数据的代表性，测量点通常部署在骨干网络的核心路由器或某个 ISP 网络的出入口处。被动测量模型主要用于测量 P2P 网络的流量大小、节点数量、连接持续时间等宏观特性。由于被动测量的前提是对 P2P 流量的准确识别，因此被动测量方法与 P2P 流量识别技术紧密关联。关于 P2P 流量识别技术已有不少的文献发表，这里不再赘述。

下面给出一个被动测量模型及实现方案。

1．模型框架

该模型根据 P2P 特定信息传播特点，采用载荷校验技术，通过对 P2P 数据包中的载荷信息进行组装、运算和校验，判断 P2P 数据包中传输的信息是否属于被监测的 P2P 特定信息，并获得相关的受众信息。该模型由样本文件生成器、监测与封堵策略编辑器、数据包捕获、信令匹配、数据包组装以及载荷校验等部分组成。模型框架如图 3-5 所示。

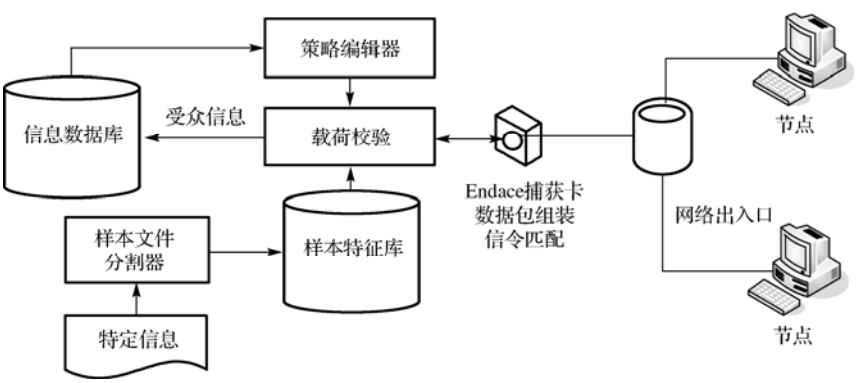


图 3-5　被动监测模型框架

被动测量模型的工作原理是：首先获取特定信息原始文件，使用样本文件分割器将原始文件处理为样本文件；然后通过部署在网络出入口处的数据包捕获设备截获 P2P 数据包，并按照 P2P 协议文件传输特点组装成可用于载荷校验的文件片（Slice），并对文件片

进行运算；载荷校验模块根据策略编辑器生成的监测策略，对运算后的文件片进行校验，判断数据包中的传输信息是否为被监测的特定信息；最后对所获取的受众信息进行处理。

模型中的主要模块功能描述如下。

（1）样本文件分割器。BitTorrent 协议为了网络共享能力的最大化以及跟踪节点的资源拥有情况，将文件先分成文件块（Piece），再分成文件片（Slice）进行传输。文件块大小范围是 32KB～2MB，默认值为 256KB，而且必须为 16KB 的倍数，文件片大小为 16KB。根据 P2P 文件最小传输单位为文件片的特点，样本文件分割器将特定信息原始文件按照文件片大小进行分割，对分割后的每一文件片使用 Hash 算法进行运算，并将运算结果拼接成样本文件，保存到样本特征库中。

（2）策略编辑器。根据被监测对象及需求，设置监测规则，包括信令匹配策略和载荷校验策略，载荷校验策略通过设置特定信息列表来实现，生成 XML 格式的监测策略，并传输给载荷校验模块。

（3）数据包捕获。采用基于 Endace 高速捕获卡的数据包捕获设备对网络出入口的流量进行捕获。

（4）信令匹配。根据信令匹配规则，对捕获数据包中的信令进行匹配，符合规则的数据包直接进行处理。信令匹配主要根据 BitTorrent 协议关键字进行。

（5）数据包组装。根据 P2P 协议中最小传输单位为文件片的特点，将捕获的多个数据包组装成文件片，并对文件片进行运算，传输给载荷校验模块。

（6）载荷校验：载荷校验模块根据监测策略，将样本文件读取到内存中，并对运算后的文件片进行校验。当同一 TCP 连接中校验成功的文件片数量超过指定阈值时，可判断该 TCP 连接中的传输内容为被监测的特定信息，并获取相应的受众信息。载荷校验的关键是校验算法及其效率。

2．载荷校验算法

载荷校验算法步骤如下：

（1）创建数据结构。根据特定信息的数量 n 创建二维 Bloom Filter[3]，二维 Bloom Filter 包含 n 个一维 Bloom Filter，每个一维 Bloom Filter 的大小根据样本文件大小进行设置，数据结构如图 3-6 所示。

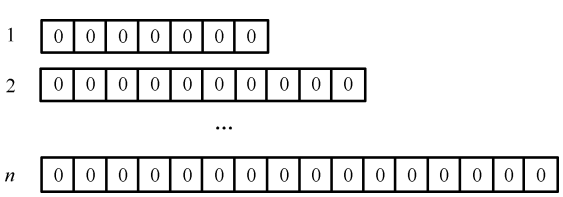


图 3-6　二维 Bloom Filter 数据结构

（2）数据映射。使用 k 个相互独立的 Hash 函数将每个样本文件中的数据映射到相应的 Bloom Filter，映射算法与标准 Bloom Filter 相同；

（3）载荷数据校验。使用 k 个相互独立的 Hash 函数分别对待检载荷进行运算，得到 k 个位置信息，使用这 k 个位置信息依次与 n 个 Bloom Filter 中位数组长度大于位置索引的 Bloom Filter 数据进行比较。与每个 Bloom Filter 中的数据比较相当于 k 次数组数据判断，效率较高。

这种二维 Bloom Filter 的载荷校验算法具有如下优点：将 $s_1+s_2+\cdots+s_n$ 个元素集合的表示压缩到 n 个长度分别为 $m_1,m_2,\cdots,m_n$ 的位数组中，大大减少了算法对内存空间的占用；将字符串依次比较算法改进为 k 次数组位数据比较算法，提高了算法执行效率；改进了标准 Bloom Filter 不能判断载荷属于哪一个特定信息的问题。

3．算法性能分析

被动测量模型主要特点是通过载荷校验对特定信息进行监测，得到与特定信息相关的受众信息。载荷校验算法是被动测量模型的关键，算法的效率决定了该模型是否可以用于实时测量，下面对算法性能进行分析。

1）误判率及相关参数分析

使用 Bloom Filter 算法时可能会出现误判，将不在该集合中的元素误判为在该集合中，这是以较小空间表示大于该空间数量的元素而产生的必然冲突，应当分析位向量长度 m 以及 Hash 函数个数 k 和误判率 R_{ec} 之间的数学关系，使得误判率 R_{ec} 尽可能小。假设 $kN<m$，且 k 个 Hash 函数是完全独立随机的。当集合 $A=\{a_1,a_2,\cdots,a_N\}$ 中的所有元素被 k 个 Hash 函数映射到长度为 m 的位数组时，这个数组中某一位还是 0 的概率为：

$$p_0=\left(1-\frac{1}{m}\right)^{kN}\approx \mathrm{e}^{-\frac{kN}{m}} \tag{3-8}$$

式中，$1/m$ 表示任意一个哈希函数选中这一位的概率。如果使用 p_0' 表示位数组中 0 的比例，则 p_0' 的数学期望为 p_0，即 $E(p_0')=p_0$。文献[4]已经证明：位数组中 0 的比例非常集中地分布在它的数学期望值附近。如果不属于 A 中的元素通过 k 次 Hash 计算后对应位置的值都为 1，此校验就是误判，可得：

$$R_{\mathrm{ec}}=(1-p_0)^k=\left(1-\left(1-\frac{1}{m}\right)^{kN}\right)^k=\mathrm{e}^{k\ln\left(1-\left(1-\frac{1}{m}\right)^{kN}\right)}\approx\left(1-\mathrm{e}^{-\frac{kN}{m}}\right)^k \tag{3-9}$$

令 $g_{\mathrm{ec}}=k\ln\left(1-\left(1-\frac{1}{m}\right)^{kN}\right)$，当 g_{ec} 取得最小值时，误判率 R_{ec} 也就取得了最小值。对 g_{ec} 进行形式变换，可得：

$$g_{\mathrm{ec}} = k\ln\left(1-\left(1-\frac{1}{m}\right)^{kN}\right) = \frac{1}{N\ln\left(1-\frac{1}{m}\right)}\ln(p_0)\ln(1-p_0) \tag{3-10}$$

根据对称性法则可知，当 $p_0 = 1/2$ 时，也就是 $k = \ln 2\cdot(m/N)$，g_{ec} 取得最小值，误判率 R_{ec} 最小，即：

$$R_{\mathrm{ec}}\big|_{\min}(1-p_0)^{\ln 2\cdot\frac{m}{N}} \approx (1-\mathrm{e}^{-\ln 2})^{\ln 2\cdot\frac{m}{N}} = \left(\frac{1}{2}\right)^{k} \approx (0.6185)^{\frac{m}{N}} \tag{3-11}$$

在确定了集合元素数 N、位数组长度 m 和 Hash 函数个数 k 后，可以得到算法误判率 R_{ec}。二维 Bloom Filter 由 n 个独立一维 Bloom Filter 组成，它的误判率 R_{ecA} 为所有一维 Bloom Filter 误判率 R_{eci} 的最大值，即：

$$\begin{aligned} R_{\mathrm{ecA}} &= \max_{i=1}^{n}(R_{\mathrm{eci}}) = \max_{i=1}^{n}(1-p_0)^k = \max_{i=1}^{n}\left(1-\left(1-\frac{1}{m_i}\right)^{ks_i}\right)^k \\ &= \max_{i=1}^{n}\left(\mathrm{e}^{k\ln\left(1-\left(1-\frac{1}{m_i}\right)^{ks_i}\right)}\right) = \max_{i=1}^{n}\left(1-\mathrm{e}^{-\frac{ks_i}{m_i}}\right)^k \end{aligned} \tag{3-12}$$

式中，R_{eci} 表示第 i 个 Bloom Filter 的误判率，s_i 表示第 i 个特定信息的文件片样本个数，m_i 表示第 i 个位数组长度。

对于标准 Bloom Filter，在已知误判率 R_{ec} 和集合元素数 N 的情况下，需要计算位数组长度 m 应当满足的条件。假定待校验集合的元素总数量为 N_{nc}，对于集合 $A=\{a_1,a_2,\cdots,a_N\}$ 中的任意一个元素，在位数组中查询时都能得到肯定结果，同时，Bloom Filter 存在一定误判率，位数组不仅能够接受集合 A 中的元素，而且还能够接受 $R_{\mathrm{ec}}(N_{\mathrm{nc}}-N)$ 个误判数据。因此，对于一个确定的位数组来说，它能够接受总共 $N+R_{\mathrm{ec}}(N_{\mathrm{nc}}-N)$ 个元素，正确数据为 N 个，所以一个确定的位数组可以表示的集合个数为：

$$\binom{N+R_{\mathrm{ec}}(N_{\mathrm{nc}}-N)}{N}$$

m 位的位数组可以有 2^m 个不同组合。因此，可表示的集合数量为：

$$2^n\binom{N+R_{\mathrm{ec}}(N_{\mathrm{nc}}-N)}{N}$$

全部待检集合中 N 个元素的集合数量为：

$$\binom{N_{\mathrm{nc}}}{N}$$

因此，要让 m 位的位数组能够表示所有 N 个元素的集合，必须满足：

$$2^n\begin{pmatrix} N+R_{\mathrm{ec}}(N_{\mathrm{nc}}-N) \\ N \end{pmatrix} \geqslant \begin{pmatrix} N_{\mathrm{nc}} \\ N \end{pmatrix}$$

对上式进行计算，可得：

$$m \geqslant \ln \frac{\begin{pmatrix} N_{\mathrm{nc}} \\ N \end{pmatrix}}{\begin{pmatrix} N+R_{\mathrm{ec}}(N_{\mathrm{nc}}-N) \\ N \end{pmatrix}} \approx \ln\left(\frac{\begin{pmatrix} N_{\mathrm{nc}} \\ N \end{pmatrix}}{\begin{pmatrix} R_{\mathrm{ec}}N_{\mathrm{nc}} \\ N \end{pmatrix}}\right) \geqslant N\frac{-k}{\ln\left(1-\mathrm{e}^{\frac{\ln R_{\mathrm{ec}}}{k}}\right)} \tag{3-13}$$

式（3-13）可以作为位数组长度 m 的估计依据。在二维 Bloom Filter 中，每个一维 Bloom Filter 的位数组长度由相应的样本文件大小决定，与其他样本文件大小和其他位数组长度无关，即：

$$m_i = s_i \frac{-k}{\ln\left(1-\mathrm{e}^{\frac{\ln R_{\mathrm{ec}}}{k}}\right)} \tag{3-14}$$

载荷校验算法对特定信息的判断依据是：当 TCP 连接中校验成功的文件片数量超过指定阈值 φ^d 时，才能确定该 TCP 连接中的传输内容为被监测或被封堵特定信息，因此，单一载荷的校验误判对整体特定信息的监测影响较小，该算法可以容忍一定的误判率。而误判率上限的提高，可以减少位数组的长度以及 Hash 函数的个数，降低对内存空间的占用，提高算法执行效率。阈值 φ^d 的取值范围一般为 5～10。

图 3-7 和图 3-8 显示了当监测 1000 个特定信息时，Hash 函数数量从 3 变化到 10 时，误判率与占用内存的变化情况。

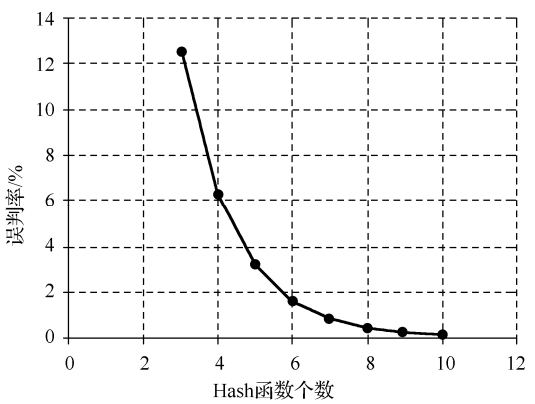


图 3-7　Hash 函数个数与误判率的关系

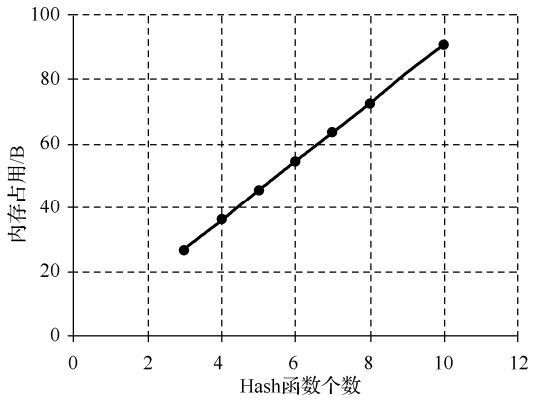


图 3-8　Hash 函数个数与占用内存的关系

从图 3-7 和图 3-8 可以看出，随着 Hash 函数个数的增加，误判率为指数降低，占用内存为线性增长。因此，Hash 函数个数设置需要在占用内存空间与误判率之间进行平衡。该算法对误判率有一定容忍度，设置 Hash 函数个数为 5，可以看到单个载荷校验的误判率为 3.12%，当阈值 φ^d 设置为 5 时，对特定信息的误判率将降低为 2.43×10^{-7}，这是一个非常小的误判率。

2）时间性能分析

载荷校验算法的时间性能包含元素映射时间和载荷校验时间。在对元素进行映射时，所需时间主要为使用 k 个 Hash 函数进行运算的时间，时间复杂度为 $O(k)$，与标准 Bloom Filter 算法所需时间是相同的。

在进行载荷校验时，算法需要进行 k 次 Hash 函数运算和最多 nk 次数组位比较运算。使用标准 Bloom Filter 的校验算法所需时间为 k 次 Hash 函数运算时间和 k 次数组位比较时间，使用二维 Bloom Filter 需要多花费最多 $(n-1)k$ 次数组位比较时间，但是可以准确知道载荷属于哪一个特定信息。假定网络出入口带宽为 1Gb/s，峰值时可以达到 128MB/s 的数据传输量，相当于要进行 128MB/16KB=8000 次比较，当 $m=1000000$、$m=1000$、$k=5$、载荷校验次数为 10000 时，载荷校验最多需要多进行 10000× (1000–1)×5 次数组位比较，所需时间平均为 0.45 秒，这样的运算效率是能够满足实时监测需求的。而且随着校验算法的进行，如果 TCP 中的传输内容被判定为被监测的特定信息，该 TCP 中后续载荷是不需要被校验的。图 3-9 显示了多种校验算法的花费时间与实时校验的要求时间对比。

从图 3-9 可以看出，使用二维 Bloom Filter 的载荷校验算法可以满足实时测量要求，而现有校验算法需要进行多次字符串比较操作，因此难以达到实时测量要求。

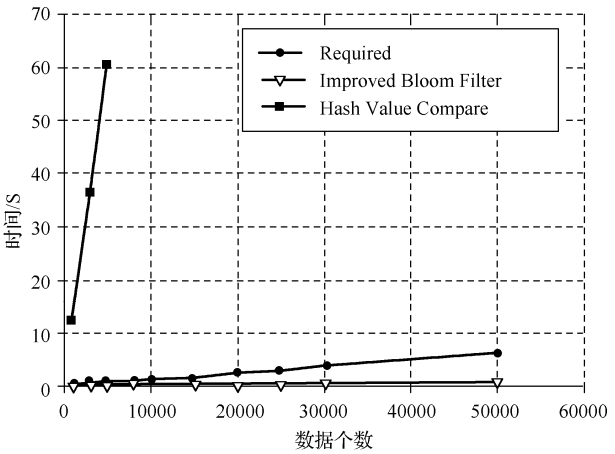


图 3-9　校验算法时间性能对比

3）空间性能分析

现有校验算法需要占用内存空间为$(s_1+s_2+\cdots+s_n)\times 160$ bit，而使用二维 Bloom Filter 的载荷校验算法将内存占用空间压缩为$(m_1+m_2+\cdots+m_n)$bit，m_i 与 s_i 之间的关系如公式（3-14）所示。设置 $k=5$，相应的 $R_{ec}=0.0312$，可得 $m_i=7.213s_i$。从而可知，使用二维 Bloom Filter 的载荷校验算法将占用的内存空间减少了 95.49%。同样以监测 1000 个平均大小为 800 MB 的特定信息为例，该算法所需的内存空间为 45.09MB，这是非常容易满足的，而且有很大的扩展空间。

根据以上对载荷校验算法的各项性能分析可知，该算法执行效率较高，占用内存较少，使用特定信息判断阈值 φ^d 后误判率较低。

3.2.3　覆盖率估计方法

主动测量模型和被动测量模型是从不同角度对 P2P 特定信息传播网络进行测量的，它们都有各自优点，也都存在相应的缺点和不足。主动测量模型可以快速获取网络中的受众信息，但是难以获取受众信息的真实覆盖率；被动测量模型可以对某一局部网络进行测量，获取该局部网络中一段时间内较为完整的受众信息，但是难以获取与该局部网络没有关联的受众信息。为了能够估计已获取受众信息的真实覆盖率，需要将主动测量模型与被动测量模型的测量结果结合起来考虑，通过对它们所获取的受众信息进行分析，得到已获取受众信息的真实覆盖率。

假定与特定信息相关的所有受众信息集合为 A_{EP}，数量为 N_{EP}，其中属于被动测量模型监测的局部网络受众信息集合为 A_{EPI}，数量为 N_{EPI}；由主动测量模型获取的与特定信息相关的受众信息集合为 N_{AP}，数量为 N_{AP}，其中包含由被动测量模型所测量的局部网络受

众信息集合为 N_{API}，数量为 N_{API}；被动测量模型所获取的与特定信息相关的受众信息集合为 A_{DP}，数量为 N_{DP}，其中属于它所监测局部网络的受众信息集合为 A_{DPI}，数量为 N_{DPI}，如图 3-10 所示。

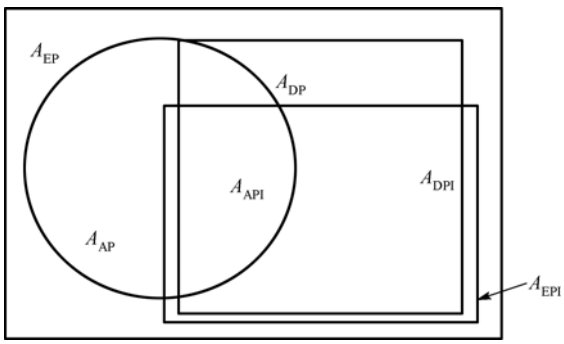


图 3-10　受众信息集合示意图

根据概率原理，所有受众信息的数量 A_{EP} 应当为：

$$\frac{N_{\mathrm{AP}}}{N_{\mathrm{EP}}}=\frac{N_{\mathrm{API}}}{N_{\mathrm{EPI}}}\Rightarrow N_{\mathrm{EP}}=N_{\mathrm{AP}}\frac{N_{\mathrm{EPI}}}{N_{\mathrm{API}}}\approx N_{\mathrm{AP}}\frac{N_{\mathrm{DPI}}}{N_{\mathrm{API}}} \tag{3-15}$$

由于被动测量模型获取的受众信息覆盖率较高，因此，$N_{\mathrm{EPI}}\approx N_{\mathrm{DPI}}$。已获取受众信息的真实覆盖率为：

$$g_{\mathrm{fE}}\approx\frac{N_{\mathrm{AP}}+(N_{\mathrm{DPI}}-N_{\mathrm{API}})+\left|A_{\mathrm{DP}}-A_{\mathrm{DPI}}\right|-\left|(A_{\mathrm{DP}}-A_{\mathrm{DPI}})\cap A_{\mathrm{AP}}\right|}{N_{\mathrm{EP}}} \tag{3-16}$$

式中，$\left|A_{\mathrm{DP}}-A_{\mathrm{DPI}}\right|$ 为集合 $A_{\mathrm{DP}}-A_{\mathrm{DPI}}$ 的元素数量。通过对 N_{EP} 和 g_{fE} 的计算，可以估计出 P2P 特定信息传播网络的规模以及已获取受众信息的真实覆盖率，有利于对特定信息传播特性进行准确分析。

3.2.4　测量方法比较

主动测量方法通过发送探测数据包和分析响应数据包来获取网络性能数据，可以获得用户感兴趣的端到端网络状况和网络行为。该方法不需要多个节点之间的协作，具有灵活方便、操作性强、可信度高、准确性好等特点。但是，主动测量方法需要相当的先验知识，而且是针对特定应用的测量，通用性较差。此外，主动测量方法还引入了额外的探测流量，增加了网络负担。在主动测量模型中，由于很难大范围地部署测量客户端，因此难以得到 P2P 网络的整体运行情况以及节点之间的连接信息。

被动测量方法属于非侵扰式的测量方法，以被动方式来收集网络流量信息，既不会增

加网络负载，也不会对节点本身造成影响，可以用于对不同 P2P 应用的测量，通用性较好。被动监测方法的主要缺点是无法深度解析 P2P 网络行为，对测量设备的要求较高。此外，被动测量方法的基础是基于数据包的特征识别，随着网络流量的高速增长以及 P2P 数据包隐蔽技术的发展，实现准确、高效的实时测量变得非常困难。

另外，目前的大多数 P2P 网络测量模型都是针对一般的 P2P 网络性能进行测量的，而本方案给出的 P2P 网络测量模型主要是针对 P2P 特定信息传播网络的拓扑特性和用户行为进行测量的，在主动测量模型中引入了 PEX 技术，不但能够获取节点之间的连接关系，而且还提高了受众信息获取效率；在被动测量模型中，采用了基于二维 Bloom Filter 的高效载荷校验算法，能够满足被动测量的实时性要求。

3.3 P2P 信息传播动力学模型

传播动力学是一种对事物传播规律进行数学建模分析的重要方法，已在多个领域取得了不少的研究成果。随着复杂网络研究的不断深入，人们发现不同事物在真实系统中的传播现象，例如，网络病毒的传播、传染病的流行、谣言的扩散等，都可以看作是在复杂网络上遵循某种规律的传播行为。运用传播动力学方法对这些事物传播过程进行建模分析，有助于揭示事物传播特性和内在规律，为推动或阻断事物传播提供理论基础。在传播动力学中，流行病传播数学模型能够很好地描述复杂网络的传播特性，是复杂网络传播动力学研究的基础。

在现有的传播动力学模型中，SEIR 模型与 P2P 特定信息传播过程比较类似。由于 P2P 特定信息传播的自身特点，SEIR 模型还难以准确地模拟 P2P 网络节点在特定信息传播过程中的状态转换，因此需要对 SEIR 模型进行改造，使之能够更准确地反映 P2P 特定信息传播过程中的节点状态转换。

下面给出一种用于描述 P2P 特定信息传播过程的传播动力学模型 SEInR，它是在 SEIR 模型的基础上改造而成的，能够比较准确地反映 P2P 特定信息传播过程中的节点状态转换。

3.3.1 SEInR 模型描述

在 P2P 文件共享系统中，用户节点既可作为文件需求者下载文件，又可作为文件供应者上传文件，流行文件的大范围传播过程与传染病传播过程相类似，可以运用传播动力学方法对 P2P 特定信息传播过程建模分析。

SEIR 模型所描述的传播过程与 P2P 特定信息传播过程比较类似，下面运用 SEIR 模型来描述 P2P 特定信息传播过程，对 SEIR 模型参数做如下的规定和描述：

（1）$S_R(t)$ 表示 P2P 网络中还没有开始特定信息下载的节点，但有可能对特定信息发

生兴趣，并进行下载。

（2）$E_R(t)$ 表示 P2P 网络中正在进行特定信息下载的节点，这些节点只有特定信息的部分内容，既下载文件内容也上传已有的文件内容。

（3）$I_R(t)$ 表示 P2P 网络中拥有完整的特定信息并进行特定信息上传的节点。

（4）$R_R(t)$ 表示 P2P 网络中针对特定信息停止共享和放弃下载的节点。

（5）$I_R(0)$ 表示 P2P 网络中初始拥有特定信息并提供下载的节点。

（6）由于离线与上线的节点数量在一段时间内基本一致，为了简化模型，本模型不考虑节点离线与上线过程。

（7）由于用户重复下载同一文件的概率极低，本模型不考虑节点的重复下载。

由于 P2P 特定信息传播的自身特点，直接使用 SEIR 模型对 P2P 特定信息传播过程进行分析时，存在如下的问题：

（1）SEIR 模型中的潜伏者是不能感染易感者的，但是在 P2P 特定信息传播过程中，潜伏节点也可以进行文件上传，相当于对易感者进行感染。

（2）SEIR 模型中的潜伏者是不能转换为治愈者的，但是在 P2P 特定信息传播过程中，部分潜伏节点，由于某种原因（如下载速度过慢或者已下载部分与预期不符等）不想继续下载，该节点可转换为治愈节点。

（3）SEIR 模型没有考虑 P2P 网络中新节点的加入和旧节点的永久退出的问题。

（4）SEIR 模型中所有感染者被治愈的概率是相同的，即对感染者不加区分。而 P2P 网络中，节点对特定信息共享意愿以及节点类型的不同，特定信息被共享的时间是不同的，也就是说感染节点被治愈的概率是不同的。

（5）在 SEIR 模型中，易感者转化为潜伏者与节点的度值有关，但是在 P2P 特定信息传播时，易感节点转化为潜伏节点与节点的度值无关，只与 P2P 网络的节点列表返回数量以及返回节点可用性有关。

针对 SEIR 模型在描述 P2P 特定信息传播过程时存在的不足，需要根据 P2P 特定信息传播特点对 SEIR 模型进行改造，主要改造内容如下：

（1）建立潜伏者对易感者的感染机制，使得易感者被感染时不但要考虑感染者因素，还需要考虑潜伏者因素。

（2）考虑潜伏者以概率 p_{ER} 直接转换为治愈者的情况。

（3）考虑 P2P 网络中新节点加入和旧节点退出，新加入节点都为易感节点。为了简化模型，将永久退出概率和新节点加入概率设为 p_F。

（4）根据节点对特定信息共享时间的不同，将感染者划分为 n_I 个子类：$I_1, I_2, \cdots, I_{n_1}$，并给每个子类设置不同的被治愈概率：$\delta_1, \delta_2, \cdots, \delta_{n_1}$。

（5）考虑 P2P 网络的节点列表返回率及节点可用率对传播模型的影响，c_u 为节点列

表返回数量、节点可用性以及实际连接率的综合表示。

改进后的模型称为 SEInR（Susceptible-Exposed-n Infected-Removed）模型，SEInR 模型的传播过程如图 3-11 所示。

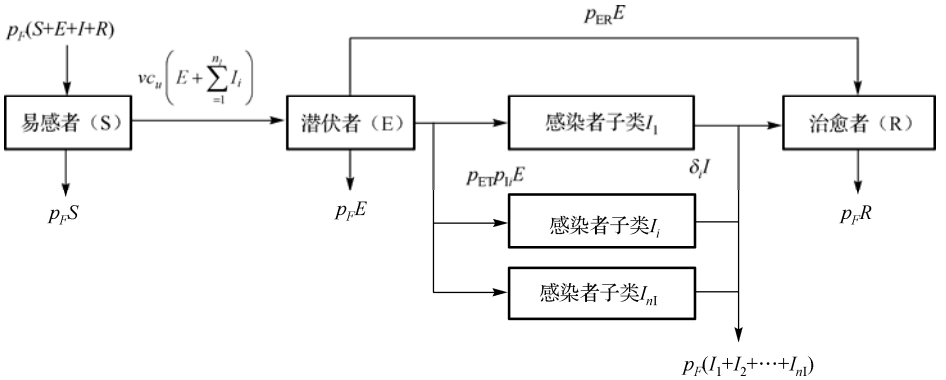


图 3-11　SEInR 模型传播过程

使用 SEInR 模型对 P2P 特定信息传播过程描述如下。

（1）易感节点发送文件请求，获取可用节点列表，可用节点列表既包含潜伏节点也包含感染节点。

（2）易感节点与可用节点列表建立连接并获取文件信息，易感节点变为潜伏节点，潜伏节点既下载文件信息也上传文件信息。

（3）潜伏节点在下载过程中以概率 p_{ER} 对特定信息失去下载需求，变为治愈节点。

（4）潜伏节点以概率 p_{EI} 下载完成后变为感染节点。

（5）感染节点对特定信息进行共享，共享完成后，感染节点变为治愈节点。

每个感染者子类根据该类平均共享时间计算该类的被治愈概率 $\delta_i(1\leqslant i\leqslant n_I)$，$n_I$ 表示感染者子类数量。单位时间内新加入节点率为常数 p_F，数量为 $p_F N_{\mathrm{B}}(t)$。每个状态中的节点都会以概率 p_F 永久退出 P2P 网络。

根据以上对 P2P 特定信息传播过程的描述，SEInR 模型的动力学方程组为：

$$\begin{cases}\dfrac{\mathrm{d}S_R}{\mathrm{d}t}=p_F-vc_uS_R\left(E_R+\sum\limits_{i=1}^{n_I}I_{\mathrm{R}i}\right)-p_FS_R\\[2ex]\dfrac{\mathrm{d}E_R}{\mathrm{d}t}=vc_uS_R\left(E_R+\sum\limits_{i=1}^{n_I}I_{\mathrm{R}i}\right)-(p_{\mathrm{EI}}+p_{\mathrm{ER}}+p_F)E_R\\[2ex]\dfrac{\mathrm{d}I_{\mathrm{R}i}}{\mathrm{d}t}=p_{\mathrm{EI}}p_{1i}E_R-(\delta_i+p_F)I_{\mathrm{R}i}\\[2ex]\dfrac{\mathrm{d}R_R}{\mathrm{d}t}=p_{\mathrm{ER}}E_R+\sum\limits_{i=1}^{n_I}\delta_iI_{\mathrm{R}i}-p_FR_R\end{cases}\tag{3-17}$$

式中，$I_{\mathrm{R}i}$ 表示感染者子类 I_i 中的节点数量占所有节点的比例，$p_{\mathrm{I}i}$ 表示新感染节点属于感染者子类 I_i 的概率，δ_i 表示感染者子类 I_i 中的节点被治愈概率，$1\leqslant i\leqslant n_I$。

3.3.2　SEInR 模型传播行为分析

对于复杂网络的传播动力学模型，主要从传播的最终稳态与动态过程两个方面进行研究。下面对 SEInR 模型的传播行为及其特性进行分析。

对 SEInR 模型的动力学方程组进行分析，可以看出总节点 $N_B(t)$ 满足：

$$\frac{\mathrm{d}N_B(t)}{\mathrm{d}t}=p_F N_B(t)-p_F(S(t)+E(t)+I(t)+R(t))=0 \tag{3-18}$$

由于在易感节点、潜伏节点、感染节点的变化过程中没有治愈节点的变化因素，略去 SEInR 模型动力学方程组中有关治愈节点的变化方程，并不影响对整个系统的研究。简化后的动力学方程组为：

$$\begin{cases}\dfrac{\mathrm{d}S_R}{\mathrm{d}t}=p_F-vc_u S_R\left(E_R+\sum\limits_{i=1}^{n_I}I_{\mathrm{R}i}\right)-p_F S_R\\[2ex]\dfrac{\mathrm{d}E_R}{\mathrm{d}t}=vc_u S_R\left(E_R+\sum\limits_{i=1}^{n_I}I_{\mathrm{R}i}\right)-(p_{\mathrm{EI}}+p_{\mathrm{ER}}+p_F)E_R\\[2ex]\dfrac{\mathrm{d}I_{R\mathrm{i}}}{\mathrm{d}t}=p_{\mathrm{EI}}p_{\mathrm{I}i}E_R-(\delta_i+p_F)I_{R\mathrm{i}}\end{cases} \tag{3-19}$$

记集合 $D_C=\{(S_R,E_R,I_{\mathrm{R}i})\,|\,S_R\in[0,1],E_R\in[0,1],I_{\mathrm{R}i}\in[0,1],1\leqslant i\leqslant n_I\}$，显然 D_C 是简化动力学方程组的一个正向不变集，以下仅在集合 D_{C} 内对方程组进行求解。

1．基本再生数计算公式推导

基本再生数 R_0 是分析传播动力学模型的重要参数。文献[5]介绍的 R_0 计算方法为：在传染病模型中，所有个体被分类到 N_r 个仓室。令 $z=(z_1,z_2,\cdots,z_i,\cdots,z_{N_r})$，$z_i$ 表示第 i 个仓室的个体数量并且 $z_i\geqslant 0$，为了分析方便，有感染者的仓室排在前面，没有感染者的仓室排在后面。令 $\tilde{F}_i(z)$ 表示仓室 i 中出现新感染者的速率（只包括新增加感染个体，不包括仓室之间个体的移动），$\tilde{V}_i^{+}(z)$ 表示其他方式进入仓室 i 的速率，$\tilde{V}_i^{-}(z)$ 表示移出仓室 i 的速率，且 $\tilde{V}_i(z)=\tilde{V}_i^{-}(z)-\tilde{V}_i^{+}(z)$。假设每个变量都是连续可微分的，那么疾病传播模型可以用下面的方程系统来表示：

$$\frac{\mathrm{d}z_i}{\mathrm{d}t}=f_{\mathrm{r}i}(z)=\tilde{V}_i^{-}(z)-\tilde{V}_i^{+}(z) \tag{3-20}$$

式中，$f_{\mathrm{r}i}(z)$ 表示第 i 个仓室中的个体数量变化方程。令 $Df_r(z_0)$ 表示 $f_{\mathrm{r}i}(z)$ 在点 z_0 处

的 Jacobian 矩阵。如果 $\tilde{F}_i(z)=0,1\leqslant i\leqslant N_r$ 时， $Df_r(z_0)$ 有负的实部。若 E_0 是模型的无病平衡点，那么 $D\tilde{F}(E_0)$ 和 $D\tilde{V}(E_0)$ 可被分别划分为：

$$D\tilde{F}(E_0)=\begin{bmatrix} F & 0 \\ 0 & 0 \end{bmatrix} 和 D\tilde{V}(E_0)=\begin{bmatrix} V & 0 \\ J_3 & J_4 \end{bmatrix}$$

$\mathbf{FV}^{-1}$ 为再生矩阵，基本再生数 R_0 是 $\mathbf{FV}^{-1}$ 的谱半径长度，即 $R_0=p(\mathbf{FV}^{-1})$。

按照上述计算基本再生数 R_0 的方法，首先将仓室按照节点感染顺序进行排列，排列后的仓室顺序为： E,I_i,S。按照动力学方程组求出 $\tilde{F}$ 和 $\tilde{V}$ 为：

$$\tilde{F}=[\tilde{F}_1(z),\tilde{F}_2(z),\cdots,\tilde{F}_{\mathrm{n}i+2}(z)]^{\mathrm{T}}=\begin{bmatrix} vc_uS_R\left(E_R+\sum_{i=1}^{n_I}I_{\mathrm{R}i}\right) \\ 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}$$

$$\tilde{V}=[\tilde{V}_1(z),\tilde{V}_2(z),\cdots,\tilde{V}_{\mathrm{n}i+2}(z)]^{\mathrm{T}}=\begin{bmatrix} (P_{\mathrm{EI}}+P_{\mathrm{ER}}+P_F)E_R \\ (\delta_1+P_F)I_{\mathrm{R}1}-P_{\mathrm{EI}}P_{\mathrm{I}1}E_R \\ \cdots \\ (\delta_{\mathrm{n}I}+P_F)I_{\mathrm{R}1}-P_{\mathrm{EI}}P_{\mathrm{In}i}E_R \\ vc_uS_R\left(E_R+\sum_{i=1}^{n_I}I_{\mathrm{R}i}\right)+P_FS_R-P_F \end{bmatrix}$$

由于系统处于平衡点时，所有仓室中的节点数量变化率为 0，即： $\mathrm{d}S_R/\mathrm{d}t=0$； $\mathrm{d}E_R/\mathrm{d}t=0$； $\mathrm{d}I_{\mathrm{R}i}/\mathrm{d}t=0$。容易发现存在无病平衡点 $E_0=(S_R=1,E_R=0,I_R=0)$。根据 Jacobian 矩阵计算公式，求出 $D\tilde{F}(E_0)$ 和 $D\tilde{V}(E_0)$ 的值为：

$$D\tilde{F}(E_0)=\begin{bmatrix} vc_u & vc_u & \cdots & vc_u & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$D\tilde{V}(E_0)=\begin{bmatrix} P_{\mathrm{EI}}+P_{\mathrm{ER}}+P_F & 0 & \cdots & 0 & 0 \\ -P_{\mathrm{EI}}P_{\mathrm{I}1} & \delta_1+P_F & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -P_{\mathrm{EI}}P_{\mathrm{In}i} & 0 & \cdots & \delta_{\mathrm{n}i}+P_F & 0 \\ vc_u & vc_u & \cdots & vc_u & P_F \end{bmatrix}$$

从 $D\tilde{F}(E_0)$ 和 $D\tilde{V}(E_0)$ 中提取出 **F** 和 **V**。令 $\omega_E = p_{\mathrm{EI}} + p_{\mathrm{ER}} + p_F$ 表示潜伏节点的减少概率，$p_{\mathrm{EI}i} = p_{\mathrm{EI}}p_{\mathrm{I}i}$ 表示潜伏节点进入感染者子类 I_i 的概率，$\omega_{\mathrm{I}i} = \delta_i + p_F$ 表示感染者子类 I_i 中的节点减少概率，计算 $\mathbf{FV}^{-1}$ 为：

$$\boldsymbol{F} = \begin{pmatrix} vc_u & vc_u & \cdots & vc_u \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad \boldsymbol{V} = \begin{pmatrix} p_{\mathrm{EI}} + p_{\mathrm{ER}} + p_F & 0 & \cdots & 0 \\ -p_{\mathrm{EI}}p_{\mathrm{I1}} & \delta_1 + p_F & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ -p_{\mathrm{EI}}p_{\mathrm{In}_I} & 0 & \cdots & \delta_{n_I} + p_F \end{pmatrix}$$

$$\mathbf{FV}^{-1} = \begin{pmatrix} vc_u & vc_u & \cdots & vc_u \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \omega_E & 0 & \cdots & 0 \\ -p_{\mathrm{EI1}} & \omega_{\mathrm{I1}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ -p_{\mathrm{EIn}_I} & 0 & \cdots & \omega_{\mathrm{In}_I} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} vc_u & vc_u & \cdots & vc_u \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \dfrac{1}{\omega_E} & 0 & \cdots & 0 \\ \dfrac{p_{\mathrm{EI1}}}{\omega_{\mathrm{E}}\omega_{\mathrm{I1}}} & \dfrac{1}{\omega_{\mathrm{I1}}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{p_{\mathrm{EIn}_I}}{\omega_{\mathrm{E}}\omega_{\mathrm{In}_I}} & 0 & \cdots & \dfrac{1}{\omega_{\mathrm{In}_I}} \end{pmatrix} = \begin{pmatrix} \dfrac{vc_u}{\omega_E}\left(1 + \sum\limits_{i=1}^{n_I} \dfrac{p_{\mathrm{EI}i}}{\omega_{\mathrm{I}i}}\right) & \dfrac{vc_u}{\omega_{\mathrm{I1}}} & \cdots & \dfrac{vc_u}{\omega_{\mathrm{In}_I}} \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

计算 $\mathbf{FV}^{-1}$ 的谱半径长度即可求得基本再生数 R_0 为：

$$R_0 = \frac{vc_u}{\omega_E}\left(1 + \sum_{i=1}^{n_I} \frac{p_{\mathrm{EI}i}}{\omega_{\mathrm{I}i}}\right) = \frac{vc_u}{(p_{\mathrm{EI}} + p_{\mathrm{ER}} + p_F)}\left(1 + \sum_{i=1}^{n_I} \frac{p_{\mathrm{EI}} + p_{\mathrm{I}i}}{(\delta_i + p_F)}\right) \tag{3-21}$$

2．平衡点存在性

系统处于平衡点时，所有仓室中的节点数量变化率为 0，容易发现模型存在无病平衡点 $E_0 = (S_R = 1, E_R = 0, I_R = 0)$。将公式（3-19）中的 S_R 和 $I_{\mathrm{R}i}$ 用 E_R 表示，可得：

$$\begin{cases} S_R = 1 - \dfrac{(p_{\mathrm{EI}} + p_{\mathrm{ER}} + p_F)}{p_F} E_R = 1 - \dfrac{\omega_E}{p_F} E_R \\ I_{\mathrm{R}i} = \dfrac{p_{\mathrm{EI}}p_{\mathrm{I}i}E_R}{(\delta_i + p_F)} = \dfrac{p_{\mathrm{EI}i}}{\omega_{\mathrm{I}i}} E_R \end{cases} \tag{3-22}$$

将公式（3-22）代入 $\mathrm{d}S_R / \mathrm{d}t = 0$；$\mathrm{d}E_R / \mathrm{d}t = 0$；$\mathrm{d}I_{\mathrm{R}i} / \mathrm{d}t = 0$，计算后可得：

$$\begin{cases} E_R^* = \left(1-\dfrac{\omega_{\mathrm{E}}}{vc_u\left(1+\sum\limits_{i=1}^{n_I}\dfrac{p_{\mathrm{EI}i}}{\omega_{\mathrm{I}i}}\right)}\right)\dfrac{p_F}{\omega_E} = \left(1-\dfrac{1}{R_0}\right)\dfrac{p_F}{\omega_E} \\ S_R^* = 1-\dfrac{\omega_E}{p_F}E_R = 1-\dfrac{\omega_E}{p_F}\left(1-\dfrac{1}{R_0}\right)\dfrac{p_F}{\omega_E} = \dfrac{1}{R_0} \\ I_{\mathrm{R}i}^* = \dfrac{p_{\mathrm{EI}i}}{\omega_{\mathrm{I}i}}E_R = \dfrac{p_{\mathrm{EI}i}p_F}{\omega_{\mathrm{I}i}\omega_E}\left(1-\dfrac{1}{R_0}\right) \end{cases} \tag{3-23}$$

当 $R_0>1$ 时，$I_{\mathrm{R}i}>0$，相应的 $S_R>0$，$E_R>0$，$R_R>0$，表示模型存在有病平衡点 $E^*=(S_R^*,E_R^*,I_R^*)$。

定理 3-1：SEInR 模型总有无病平衡点 $E_0=(S_R=1,E_R=0,I_R=0)$，当 $R_0>1$ 时，除存在无病平衡点外，还存在有病平衡点 $E^*=(S_R^*,E_R^*,I_R^*)$。

3．模型稳定性

下面对平衡点稳定性以及传播网络的全局稳定性进行分析。

定理 3-2：当 $R_0\leqslant 1$ 时，SEInR 模型的无病平衡点 $E_0=(S_R=1,E_R=0,I_R=0)$ 是局部渐进稳定的；当 $R_0>1$ 时，E_0 是不稳定的。

证明：对模型动力学方程组求 Jacobian 矩阵 $\boldsymbol{J}_{\mathrm{SEInR}}$ 及其在 E_0 处的值为：

$$\boldsymbol{J}_{\mathrm{SEInR}}=\begin{pmatrix} -vc_u\left(E_R+\sum\limits_{i=1}^{n_I}I_{\mathrm{R}i}\right)-p_F & -vc_uS_R & -vc_uS_R & \cdots & -vc_uS_R \\ vc_u\left(E_R+\sum\limits_{i=1}^{n_I}I_{\mathrm{R}i}\right) & vc_uS_R-(p_{\mathrm{EI}}+p_{\mathrm{ER}}+p_F) & vc_uS_R & \cdots & vc_uS_R \\ 0 & p_{\mathrm{EI}}p_{\mathrm{I1}} & -(\delta_1+p_F) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & p_{\mathrm{EI}}p_{\mathrm{In}_I} & 0 & \cdots & -(\delta_{n_I}+p_F) \end{pmatrix}$$

$$\boldsymbol{J}_{\mathrm{SEInR}}(E_0)=\begin{pmatrix} -p_F & -vc_u & -vc_u & \cdots & -vc_u \\ 0 & vc_u-(p_{\mathrm{EI}}+p_{\mathrm{ER}}+p_F) & vc_u & \cdots & vc_u \\ 0 & p_{EI}p_{\mathrm{I1}} & -(\delta_1+p_F) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & p_{\mathrm{EI}}p_{\mathrm{In}_I} & 0 & \cdots & -(\delta_{n_I}+p_F) \end{pmatrix}$$

其中一个特征根为 $-p_F$，另外 2 个特征根为以下矩阵的特征根。

$$\begin{pmatrix} vc_u-(p_{\mathrm{EI}}+p_{\mathrm{ER}}+p_F) & vc_u & \cdots & vc_u \\ p_{\mathrm{EI}}p_{\mathrm{I1}} & -(\delta_1+p_F) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ p_{\mathrm{EI}}p_{\mathrm{In}_I} & 0 & \cdots & -(\delta_{n_I}+p_F) \end{pmatrix}$$

$$=\begin{pmatrix} vc_u-\omega_E & vc_u & \cdots & vc_u \\ p_{\mathrm{EI1}} & -\omega_{\mathrm{I1}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ p_{\mathrm{EIn}_I} & 0 & \cdots & -\omega_{\mathrm{In}_I} \end{pmatrix}=\begin{pmatrix} \omega_E(R_0-1) & vc_u & \cdots & vc_u \\ 0 & -\omega_{I1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -\omega_{\mathrm{In}_I} \end{pmatrix}$$

当 $R_0 \leqslant 1$ 时，该矩阵的所有特征根均不存在正实部，因此，E_0 是局部稳定的；当 $R_0>1$ 时，该矩阵至少有一个具有正实部的特征根，因此，E_0 是不稳定的。

定理 3-3：当 $R_0 \leqslant 1$ 时，SEInR 模型的无病平衡点 E_0 是全局渐进稳定的。

证明：构造 Liapunov 函数 $V_R(t)$，并计算其导数，可得：

$$V_R(t)=p_{\mathrm{EI}}E_R+(p_{\mathrm{EI}}+p_{\mathrm{ER}}+p_F)\sum_{i=1}^{n_I}I_{\mathrm{R}i} \tag{3-24}$$

$$\frac{\mathrm{d}V_R(t)}{\mathrm{d}t}=p_{\mathrm{EI}}\frac{\mathrm{d}E_R}{\mathrm{d}t}+(p_{\mathrm{EI}}+p_{\mathrm{ER}}+p_F)\sum_{i=1}^{n_I}\frac{I_{\mathrm{R}i}}{\mathrm{d}t}=\omega_{\mathrm{E}}p_{\mathrm{EI}}E_R(R_0S_R-1) \tag{3-25}$$

由于 $0 \leqslant S_R \leqslant 1$，所以当 $R_0 \leqslant 1$ 时，$\mathrm{d}V_R(t)/\mathrm{d}t \leqslant 0$，且只有在无病平衡点 $E_0=(S_R=1, E_R=0, I_R=0)$ 上，$\mathrm{d}V_R(t)/\mathrm{d}t=0$。由 Lasalle 不变性原理及极限方程理论可知，当 $R_0 \leqslant 1$ 时，SEInR 模型的无病平衡点 E_0 是全局渐进稳定的。

定理 3-4：当 $R_0>1$ 时，SEInR 模型的唯一有病平衡点 E^* 是局部渐进稳定的。

证明：将有病平衡点 E^* 代入模型的 Jacobian 矩阵并进行化简后可得：

$$\boldsymbol{J}_{\mathrm{SEInR}}(E^*)=\begin{pmatrix} -a_{11} & -S_R^*\omega_E R_0 & -vc_uS_R^* & \cdots & -vc_uS_R^* \\ 0 & \dfrac{\omega_E}{a_{11}}(p_F(R_0-1)-2R_0\omega_E E_R^*) & \dfrac{p_E}{a_{11}}vc_uS_R^* & \cdots & \dfrac{p_F}{a_{11}}vc_uS_R^* \\ 0 & 0 & -(\delta_1+p_F) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -(\delta_{n_I}+p_F) \end{pmatrix}$$

式中，$a_{11}=vc_u\left(E_R+\sum_{i=1}^{n_I}I_{\mathrm{R}i}\right)+p_F$。

当 $R_0>1$ 时，在有病平衡点 E^* 上，$E_R^*=(1-1/R_0)p_F/\omega_E>0$，为了使 E^* 是局部渐进稳定的，需要得到 $\omega_E(p_F(R_0-1)-2R_0\omega_E E_R^*)/a_{11}<0$，通过反推法证明。

$$\frac{\omega_E}{a_{11}}(p_F(R_0-1)-2R_0\omega_E E_R^*)<0 \Rightarrow p_F(R_0-1)-2R_0\omega_E E_R^*<0$$

$$\Rightarrow p_F(R_0-1)<2R_0\omega_E E_R^* \Rightarrow p_F(R_0-1)<2R_0\omega_E\left(1-\frac{1}{R_0}\right)\frac{p_F}{\omega_E}$$

$$\Rightarrow p_F(R_0-1)<2p_F(R_0-1)\Rightarrow 1<2$$

由于 1<2 总是成立的，因此，E^* 是局部渐进稳定的。

定理 3-5：当 $R_0>1$ 时，SEInR 模型的唯一有病平衡点 E^* 是全局渐进稳定的。

证明：令 $E_R=E_R^*(1+x_E)$，$I_{Ri}=I_{Ri}^*(1+y_{Ii})$，$1\leqslant i\leqslant n_I$，构造 Liapunov 函数为：

$$\begin{cases} V_R(t)=-V_0+\sum_{i=1}^{n_I}V_i \\ V_0=\dfrac{(x_E)^2}{2} \\ V_i=\dfrac{H_{ki}I_{Ri}^*}{(\delta_i+p_F)}(y_{Ii}-\ln(1+y_{Ii})),1\leqslant i\leqslant n_I \end{cases} \tag{3-26}$$

$$\frac{\mathrm{d}V_R(t)}{\mathrm{d}t}=-\left(p_E+\sum_{i=1}^{n_I}I_{Ri}^*(1+y_{Ii})\right)(x_E)^2$$

$$-E_R^*(1+x_E)\sum_{\substack{i,j=1\\ i<j}}^{n_I}\frac{v^2p_{Ii}I_{Ri}^*}{(\delta_i+p_F)(1+y_{Ii})(1+y_{Ij})}(y_{Ij}-y_{Ii})^2\leqslant 0$$

只有在有病平衡点 E^* 时，$x_E=0,y_{Ii}=0,1\leqslant i\leqslant n_I$，$\mathrm{d}V_R(t)/\mathrm{d}t=0$。由定理 3-4、Lasalle 不变性原理以及极限方程理论可知，当 $R_0>1$ 时，SEInR 模型的有病平衡点 E^* 是全局渐进稳定的。

综合以上对 SEInR 模型的分析，模型基本再生数 R_0 为：

$$R_0=\frac{vc_u}{\omega_E}\left(1+\sum_{i=1}^{n_I}\frac{p_{EIi}}{\omega_{Ii}}\right)=\frac{vc_u}{(p_{EI}+p_{ER}+p_F)}\left(1+\sum_{i=1}^{n_I}\frac{p_{EI}p_{Ii}}{(\delta_i+p_F)}\right) \tag{3-27}$$

当 $R_0\leqslant 1$ 时，模型存在无病平衡点 E_0，且 E_0 是全局渐进稳定的；当 $R_0>1$ 时，模型存在无病平衡点 E_0 和唯一的有病平衡点 E^*，其中 E_0 是不稳定的，E^* 是全局渐进稳定的。

3.3.3 SEInR 模型传播特性分析

对于 P2P 特定信息传播来说，已传播节点表示已经完成下载的节点，而只要开始下载的节点，不管其是否完成下载，这里统称为已下载节点。下面对 SEInR 模型传播特性进行分析。

1．传播速率及影响因素

这里的传播速率是指已传播节点的数量变化速率，不包含正在下载节点和中途取消节点。结合 SEInR 模型，单位时间内从潜伏节点变化为感染节点的数量为：

$$\Delta_E(t) = p_{\mathrm{EI}} E(t) \tag{3-28}$$

相应的传播节点数量变化速率为：

$$\begin{aligned} \frac{\mathrm{d}\Delta_E(t)}{\mathrm{d}t} &= p_{\mathrm{EI}} \frac{\mathrm{d}E(t)}{\mathrm{d}t} \\ &= p_{\mathrm{EI}} \left(v c_u \frac{S(t)}{N_B(t)} \left(E(t) + \sum_{i=1}^{n_I} I_i(t) \right) - (p_{\mathrm{EI}} + p_{\mathrm{ER}} + p_F) E(t) \right) \end{aligned} \tag{3-29}$$

对传播模型方程组及上式的分析可知，传播速率主要由以下因素决定：v、c_u、p_{ER}、p_{EI}、p_F、$p_{\mathrm{I}i}$、δ_i。

v 表示潜伏节点和感染节点对易感节点的感染概率，在 P2P 网络中，该指标主要与特定信息被关注程度有关，特定信息被关注程度越高，v 值越大。

c_u 是节点返回数量、节点可用性及实际连接率的综合表示，反映了 P2P 网络返回可用节点的能力以及易感节点与可用节点连接成功的概率。

p_{ER} 表示节点中途取消下载概率，该概率主要是由于下载时间过长，用户不愿等待而产生的。p_{ER} 越大，取消下载节点越多，转换为已传播节点数量越少，p_{ER} 越小，取消下载节点越少，转换为已传播节点数量越多。

p_{EI} 表示下载节点下载完成概率，该概率是特定信息大小与下载速度的复合函数。特定信息大小越大，下载所需时间越长，p_{EI} 越小；下载速度越快，下载所需时间越短，p_{EI} 越大。当 p_{EI} 越大时，下载节点完成下载并转换为已传播节点的数量越多，而剩余下载节点越少，下一单位时间内转换为已传播节点数量将会变少。当 $p_{\mathrm{EI}} \to 0$ 时，下载节点无法完成下载，不能转换为已传播节点，传播速率趋向于 0；当 $p_{\mathrm{EI}} \to 1$ 时，传播速率为已下载节点的生成速率。

p_F 表示新节点加入或旧节点永久退出 P2P 网络的概率，该概率主要是由于用户个人行为造成的，数值一般较小，对 P2P 特定信息传播影响甚微。p_F 越大，退出节点越多，转换为已传播节点数量越少。

$p_{\mathrm{I}i}$ 表示潜伏节点转换为感染节点时进入感染者子类 I_{i} 的概率，该概率主要与各个感染者子类中的节点数量相关，子类中的节点数量越多，该概率越大。

δ_i 表示感染者子类 I_i 中的节点取消特定信息共享的概率，节点对特定信息共享时间越长，δ_i 越小，传播能力越强。该概率主要与用户使用习惯有关，相关研究表明，P2P 网络上存在着大量“搭便车”现象，这种情况会导致 δ_i 变大，传播能力减弱。为了提高 P2P

网络传播能力，部分软件采取了客户积分手段，鼓励对已下载文件进行共享，减少“搭便车”现象，增加特定信息共享时间。

在 SEInR 模型中，潜伏节点和感染节点都具有传染能力。而在 P2P 文件共享系统中，返回的节点列表中既有潜伏节点也有感染节点，必须正确地区分这些节点。这里根据节点的资源拥有率作为区分依据，由于潜伏节点是正在进行下载的节点，感染节点是已经完成下载并进行上传的节点。因此，如果节点的资源拥有率没有达到 100%，则该节点为潜伏节点；否则，该节点为感染节点。

2．传播时间及传播规模

当基本再生数 $R_0 \leqslant 1$ 时，模型只有无病平衡点 E_0，且 E_0 是全局渐进稳定的。因此，P2P 特定信息传播时间 T_S 为 E_0 出现的时间点，此时：

$$S_R(t)=\int_0^{T_S}\frac{\mathrm{d}S_R(t)}{\mathrm{d}t}\mathrm{d}t=1 \tag{3-30}$$

求解公式（3-30）可得 P2P 特定信息的传播时间 T_S。当基本再生数 $R_0>1$ 时，由于有病平衡点是全局渐进稳定的，P2P 特定信息在网络中的传播始终存在，传播时间 $T_S\to\infty$。P2P 特定信息的传播规模分为两种情况：已下载节点规模和已传播节点规模。根据对已下载节点和已传播节点的定义，可得到已下载节点规模 $\mathrm{SC}_E(t)$ 和已传播节点规模 $\mathrm{SC}_I(t)$ 分别为：

$$\begin{cases}\mathrm{SC}_E(t)=\int_0^t vc_u\dfrac{S(t)}{N_B(t)}(E(t)+I(t))\mathrm{d}t\\ \mathrm{SC}_I(t)=\int_0^t p_{\mathrm{EI}}E(t)\mathrm{d}t\end{cases} \tag{3-31}$$

当基本再生数 $R_0\leqslant 1, t>T_S$ 时，传播模型已进入无病传播状态，不会产生新的已下载节点和已传播节点。因此，相应的传播规模为常数 $\mathrm{SC}_E(T_S)$ 和 $\mathrm{SC}_I(T_S)$。在其他情况下，传播模型将会不断产生新的已下载节点和已传播节点，$\mathrm{SC}_E(t)$ 和 $\mathrm{SC}_I(t)$ 是与时间 t 相关的单调增函数。

3.3.4 SEInR 模型验证

下面通过数据仿真方法对 SEInR 模型进行分析和验证。在传播动力学模型中，基本再生数 R_0 是分析模型性能的重要参数，SEInR 模型的基本再生数 R_0 见公式（3-27）。

将感染节点(I)划分为三个子类：长时间在线共享子类(I_1)、临时共享子类(I_2)和“搭便车”子类(I_3)。设定特定信息文件大小为 600MB，节点平均下载速度为 100KB/s，单位测量时间为 10 分钟，则潜伏节点平均寿命为 10 个单位时间，单位时间内从潜伏节点转换为感染节点的概率 p_{EI} 为 0.1。其他参数的默认值设置如下：$v=0.015$，$c_u=10$，$p_{\mathrm{ER}}=0.001$，

$p_F = 0.001$，$S(0) = 99999$，$I_i(0) = (1,0,0)$，$p_{\mathrm{Ii}} = (0.01, 0.39, 0.6)$，$\delta_i = (0.0001, 0.3, 0.95)$。各个状态的节点比率变化如图 3-12 所示，感染者子类中的节点比率变化如图 3-13 所示。

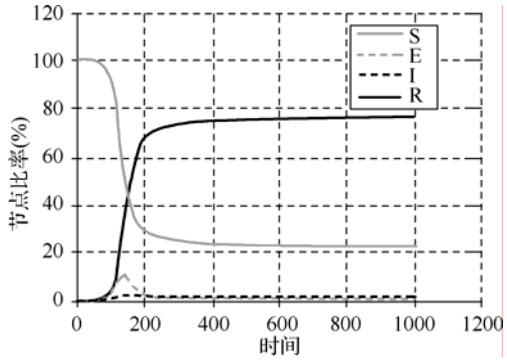


图 3-12　各个状态节点比率变化图

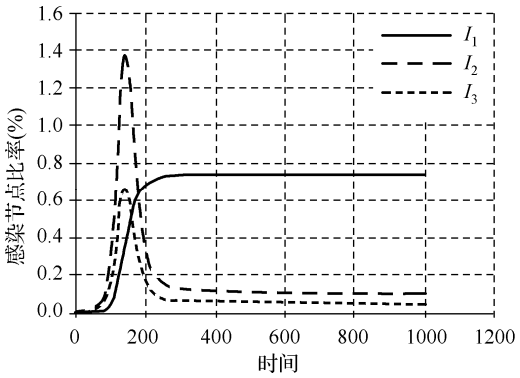
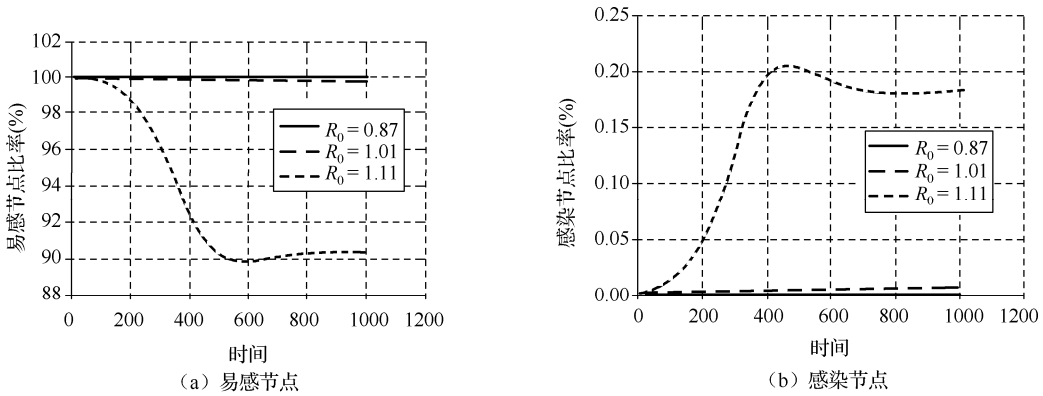


图 3-13　感染者子类节点比率变化图

从图 3-12 可以看出，$S_R(t)$ 开始下降较慢，随后有一个快速下降过程，并逐渐平缓，最终趋于平衡点中的 S_R^*；$R_R(t)$ 的变化过程与 $S_R(t)$ 相反，开始上升较慢，随后有一个快速上升过程，并逐渐平缓，最终趋于平衡点中的 R_R^*；$E_R(t)$ 与 $I_R(t)$ 的比率变化过程都是先上升后下降，最终趋于平衡点中的 E_R^* 和 I_R^*，潜伏节点的高峰值要比感染节点的高峰值大，几乎是同时到达的，高峰值大小与基本再生数 R_0 有关。

从图 3-13 可以看出，由于 I_1 中的节点长时间在线，被治愈概率和进入概率都较低，节点比率上升后下降缓慢，基本维持在高位水平 I_{R1}^*；I_2 中的节点被治愈概率和进入概率都居中，节点比率快速变化，最终稳定在平衡点中的 I_{R2}^*；虽然 I_3 中的节点进入概率很高，但是被治愈概率也很高，属于下载完成后不进行共享的“搭便车”节点，因此，I_3 中的节点比率高峰值较 I_2 中的要小，并且最终在平衡点中的值 I_{R3}^* 要小于 I_{R2}^*。

$R_0=1$是判断特定信息传播规律的重要参数，图 3-14 显示了 R_0 取不同值时，易感节点与感染节点的比率变化。从图 3-14 可以看出，当 $R_0 \leqslant 1$时，感染节点比率趋向于 0，易感节点比率趋向于 1，也就是趋向于模型的无病平衡点，并且是全局稳定的；当 $R_0>1$ 时，感染节点比率趋向于大于 0 的数值，并且是全局稳定的；当 R_0 越大时，感染节点比率越大，传播速度越快。



（a）易感节点　　（b）感染节点

图 3-14　R_0 对模型的影响

感染概率 v 代表了特定信息的传播能力，在其他参数固定不变的情况下，将 v 分别设置为：0.01、0.015、0.03，基本再生数 R_0 经过计算后分别为：2.90、4.35、8.71。SEInR 模型中的感染节点比率变化情况如图 3-15 所示，从图 3-15 可以看出，感染概率越大，感染高峰到来越快，感染高峰值越大，从感染高峰下降速率越快，到达平衡点所使用时间越短，最终平衡点的 I_R^* 越大。并且随着感染概率的增加，图形由平缓变得尖锐，高峰值与平衡点中的 I_R^* 差值越大。这些情况说明感染概率 v 较大时，特定信息传播过程变化幅度较大。

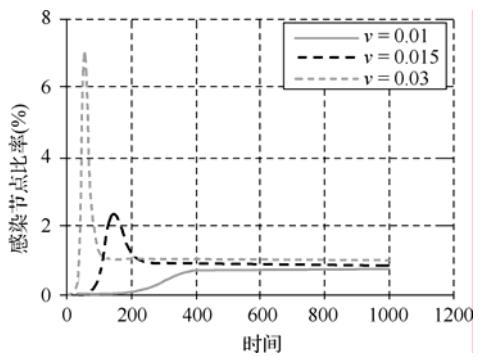


图 3-15　感染概率对模型的影响

为了分析初始感染节点对模型的影响，在其他参数固定不变的情况下，将初始感染节点数量分别设置为：1、5、20，模型中的感染节点比率变化情况如图 3-16 所示，从图 3-16 可以看出，随着初始感染节点的增加，传播过程中高峰值出现时间会提前，但是高峰值大小相同，并且高峰值出现后的传播过程是相同的。这种情况说明，初始感染节点数量的变化，只会影响感染节点比率高峰值出现时间的早晚，特定信息的整体传播过程是相同的。

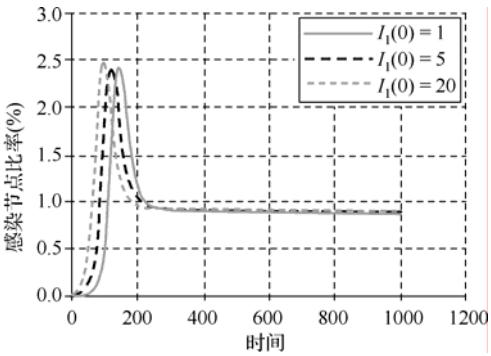


图 3-16　初始感染节点数量对模型的影响

特定信息大小与下载带宽会影响模型参数 p_{EI} 和 p_{ER}，图 3-17 显示了在固定其他参数的情况下，将 p_{EI} 分别设置为 0.2、0.1、0.01 时，p_{EI} 对模型的影响。从图 3-17 可以看出，p_{EI} 值越大，感染节点比率高峰值出现越早，平衡点中的 I_R^* 越大。但是高峰值大小与 p_{EI} 无关，平衡点出现时间早晚也与 p_{EI} 无关。说明特定信息大小只会影响高峰值出现早晚和平衡点中的 I_R^* 大小。

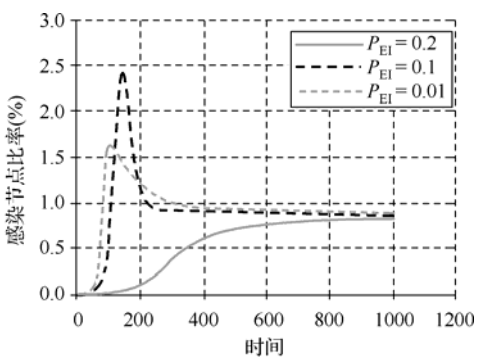


图 3-17　p_{EI} 对模型的影响

图 3-18 显示了在固定其他参数的情况下，将 p_{ER} 分别设置为 0.0001、0.001、0.01、0.1、0.5 时，p_{ER} 对模型的影响。从图 3-18 可以看出，随着 p_{ER} 增大，特定信息传播范围

越来越小，平衡点中的 I_{R}^{*} 也越来越小。特别是当 $p_{\mathrm{ER}}=0.5$ 时， $R_0=0.31$，特定信息传播趋向于无病传播。

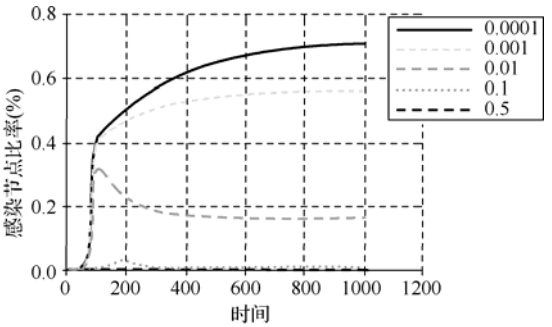


图 3-18　p_{ER} 对模型的影响

不同感染者子类有不同的治愈率和不同的进入概率，图 3-19 显示了在治愈率 δ_i 不变的情况下，进入概率 $p_{\mathrm{I}i}$ 变化时对模型的影响，治愈率设置为 $\delta_i=(0.0001,0.3,0.95)$，进入概率 $p_{\mathrm{I}i}$ 设置 3 组不同的值(0.0001、0.2999、0.7)，(0.01、0.39、0.6)，(0.1、0.5、0.4)。从图 3-19 可以看出，当进入较低治愈率子类的概率不断提高时，传播模型平衡点中的 I_R^{*} 也随之迅速提高。主要是由于该概率的提高，长时间在线节点增加，这些节点以较小概率转换为治愈节点，导致感染节点比率上升。

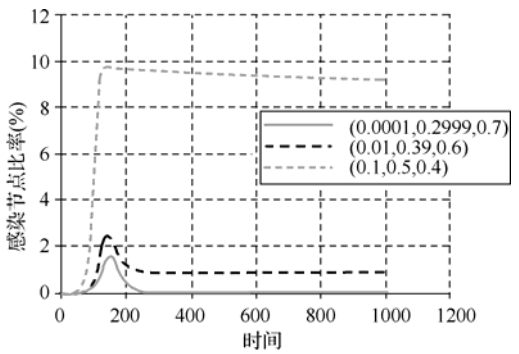


图 3-19　$p_{\mathrm{I}i}$ 对模型的影响

图 3-20 显示了在进入概率 $p_{\mathrm{I}i}$ 不变的情况下，治愈率 δ_i 变化时对模型的影响，进入概率设置为 $p_{\mathrm{I}i}=(0.01,0.39,0.6)$，治愈率 δ_i 设置 3 组不同的值(0.0001、0.3、0.95)，(0.01、0.5、0.96)，(0.1、0.7、0.99)。从图 3-20 可以看出，随着治愈率 δ_i 的提高，感染节点比率高峰值降低，平衡点中的 I_R^{*} 迅速降低，特别是当 $\delta_1\geqslant 0.01$ 时， $I_R^{*}\to 0$，模型近似于无病传播状态。说明为了提高特定信息传播范围，需要增加感染节点在线时长，特别是长期在线节点的在线时长。

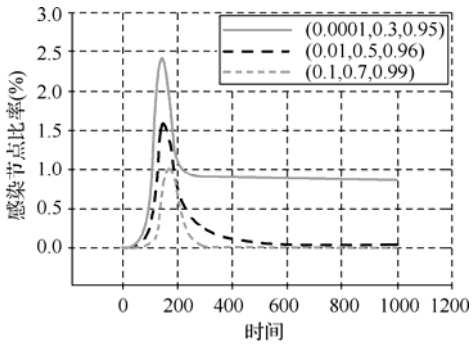


图 3-20　δ_i 对模型的影响

p_F 表示新节点加入和旧节点退出概率，该参数对模型的影响如图 3-21 所示，固定其他参数，将 p_F 分别设置为 0.01、0.001、0.0001。从图 3-21 可以看出，节点变换频率增大时，平衡点中的 I_R^* 将会增加，但是当 p_F 小于某一数值时，该参数对模型传播的影响将变小，如图 3-21 中 $p_F = 0.001$ 和 $p_F = 0.0001$ 所示。

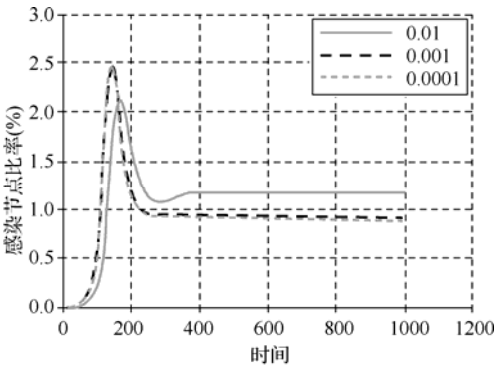


图 3-21　p_F 对模型的影响

为了验证 SEInR 模型是否能够准确描述 P2P 特定信息的传播过程，使用 NS2 仿真平台对 P2P 特定信息的传播过程进行模拟，并与 SEIR 模型和 SEInR 模型的传播过程进行比较，比较结果如图 3-22 所示。从图 3-22 可以看出，SEIR 模型的传播过程与实际情况差距较大，而 SEInR 模型能够较为准确地模拟了 P2P 特定信息的传播过程，可用于对 P2P 特定信息传播规律的分析。

综上所述，在 P2P 文件共享系统中，特定信息的大范围传播过程与传染病传播过程具有极大的相似性，传播动力学是研究 P2P 特定信息传播特性和内在规律的重要手段。SEInR 模型能够较好地描述 P2P 特定信息传播过程，为深入分析 P2P 特定信息传播特性提供了理论基础和科学依据。SEInR 模型是根据 P2P 特定信息传播特点，对 SEIR 模型进

行改造而成的，改造的主要内容包括：将传统感染者分为 n_I 个子类，每个子类根据节点在线时长赋予不同的治愈概率；考虑潜伏节点的感染能力和转换为治愈节点的概率；引入节点加入与退出机制；考虑节点列表返回能力的影响。

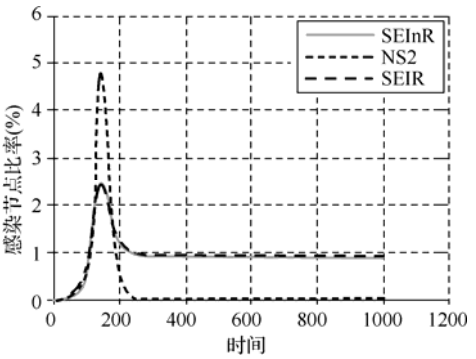


图 3-22　SEInR 模型效果验证

在 SEInR 模型基础上，推导出基本再生数 R_0 的计算公式，并对模型传播行为及其特性进行了深入分析，包括：无病平衡点及有病平衡点的存在性、模型稳定性、模型传播速率与影响因素、传播时间与传播规模等。仿真实验结果表明，SEInR 模型所描述的传播过程比较符合 P2P 特定信息传播规律，模型中的各个参数能够比较准确地描述传播过程中的各种影响因素。

3.4　P2P 特定信息传播特性

P2P 文件共享系统是建立在互联网上的应用层覆盖网络，随着特定信息的传播，自主地构成一个 P2P 特定信息传播网络，具有天然的动态特性。为了认识特定信息的传播规律，需要对 P2P 特定信息传播网络特性进行分析，这里主要分析“元信息”属性、网络拓扑特性和用户行为特性等。

在分析 P2P 特定信息传播网络特性时，其理论依据是 SEInR 模型，基本数据是通过网络测量所获取的受众信息。

3.4.1　“元信息”属性分析

“元信息”是启动 P2P 特定信息传播任务的基本信息，“元信息”属性主要有文件分类及大小、“元信息”发布规律、“元信息”流行度等，它们将对 P2P 特定信息传播过程产生影响。下面结合 SEInR 模型，对“元信息”主要属性及其对传播特性的影响进行分析，为确定 SEInR 模型中的相关参数提供依据。

1．文件分类及长度

“元信息”对应的文件长度影响着文件下载时间，而文件下载时间与 SEInR 模型中的参数 p_{EI} 和 p_{ER} 有着密切的关系，下载时间越短，p_{EI} 越大，p_{ER} 越小。如果下载时间过长，取消下载概率 p_{ER} 会大幅度提高。文件分类决定着文件长度范围，例如，高清影视的文件平均长度要大于普通影视的文件平均长度。因此，对文件分类及长度进行分析，有助于确定 SEInR 模型的相关参数。

下面通过实例对文件分类及长度进行分析。在本实例中，采用主题网络爬虫工具抓取飞鸟娱乐（http://bbs.hdbird.com/）和悠悠鸟影视（http://bbs.uuniao.com/）网站的“元信息”数据。为了分析的方便性，将“元信息”分为以下几类：高清影视、普通影视、连续剧、音乐、游戏和其他共六大类。数据抓取开始时间为 2010 年 7 月，总共历时 188 天，每天抓取 2 次。图 3-23 和图 3-24 为抓取到的所有分类“元信息”数量和文件总长度分布图。

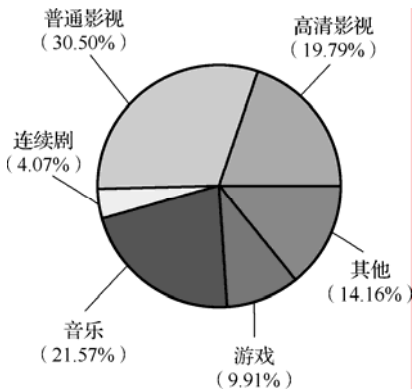


图 3-23　“元信息”数量分布图

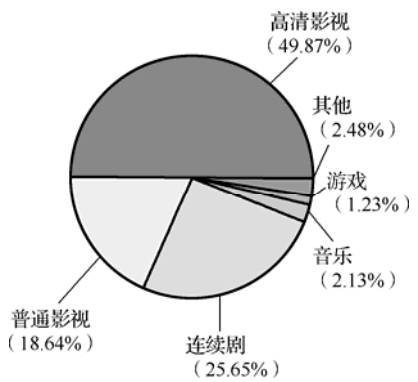


图 3-24　文件总长度分布图

从图 3-23 可以看出，获取到的“元信息”类型主要为普通影视、高清影视和音乐文件，表明在 P2P 网络上传输的文件主要以电影文件和音乐文件为主。从图 3-24 可以看出，由于高清影视单个文件较长，文件总长度所占比例较大，而对于音乐“元信息”来说，虽然其数量较多，但由于单个文件较短，文件总长度所占比例较小。通过对“元信息”数量和文件长度进行分类处理，可以得到每个分类中不同长度的文件所占比例。例如，高清影视的文件长度主要集中在 1～2GB（43.40%）以及 4～6GB（34.35%），这主要是由于高清影视文件格式分为 rmvb 格式和 mkv 格式，而且文件主要分辨率为 720P。通过对文件长度分布的分析，可以确定传播模型参数 p_{EI} 的大概取值范围。

2.“元信息”发布规律

“元信息”是由资源拥有者自愿发布的，网站只是提供发布平台，并不实际发布“元信息”。发布者既要发布“元信息”，还要作为种子节点加入 P2P 网络中，对发布的文件信息进行共享，只有当 P2P 网络中拥有该文件的节点达到一定规模后，该节点才能退出 P2P 网络，否则将导致其他节点无法完整地下载到文件信息。因此，分析“元信息”发布规律不但对设置数据获取时间间隔有重要的意义，而且对研究种子节点在线时长及传播模型的治愈率 δ_i 具有重要的参考价值。

图 3-25 显示了不同类型的“元信息”数量变化情况，每条曲线的起点是第一次完整获取后得到的“元信息”数量。从图 3-25 可以看出，“元信息”的数量变化曲线类似于一条直线，直线的斜率为“元信息”的每日平均更新速度，更新速度越快，斜率越大。

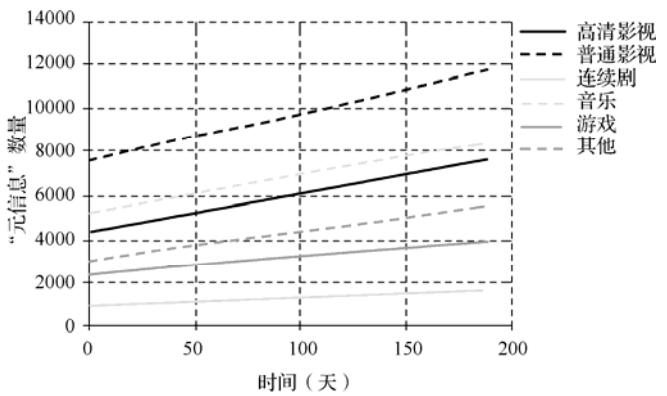


图 3-25 “元信息”数量变化图

图 3-26 显示了几种主要类型“元信息”每日更新数量的变化情况，从图 3-26 可以看出，“元信息”每日更新数量比较随机，但是有一定的周期性。高清影视平均更新速度为 18 个/天，普通影视平均更新速度为 23 个/天，连续剧平均更新速度为 4 个/天，音乐平均更新速度为 16 个/天。

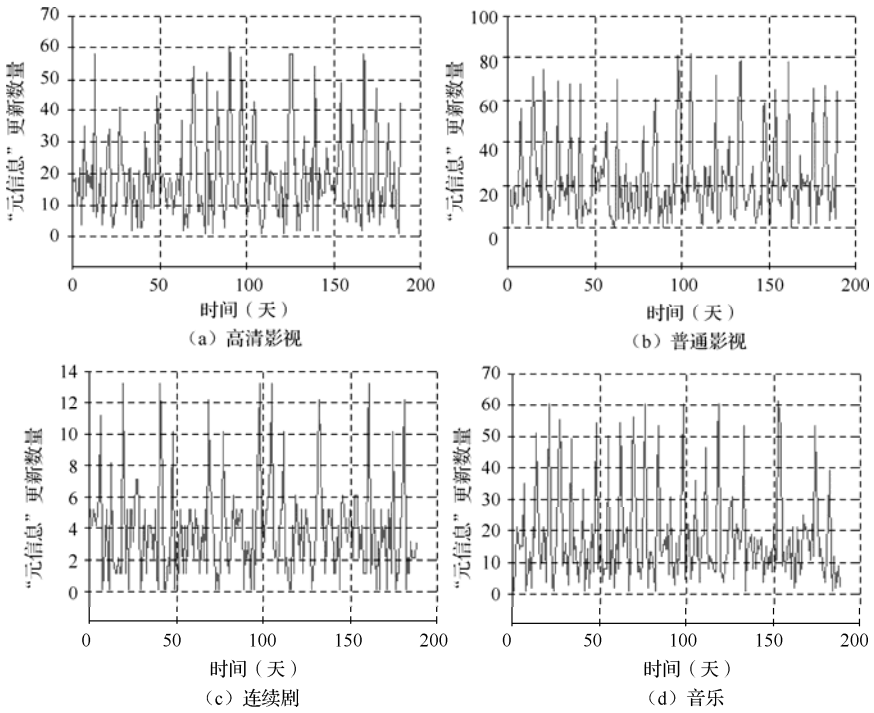


图 3-26　“元信息”更新数量变化图

下面以日和周为周期，对“元信息”更新时间进行分析。在日周期中，根据生活规律将每天分为 4 个时段：0～6 点、6～12 点、12～18 点和 18～0 点，统计每个时段更新数量，如图 3-27 所示，从图 3-27 可以看出，更新比较集中的是 18～0 点时段，而 6～12 点时段是更新最少时段，虽然，0～6 点的更新数量较多，但是其中 2～6 点时段内的更新数量极少，所占比例只有 0.189%。在周周期中，按照每周 7 天进行分段统计，如图 3-28 所示，从图 3-28 可以看出，周六和周日的更新速度明显要高于其他天的更新速度。

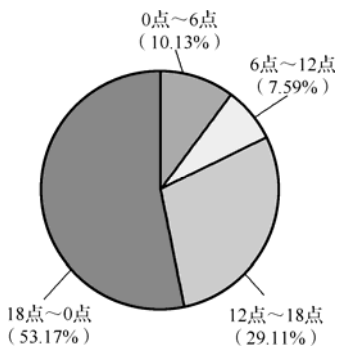


图 3-27　“元信息”更新日周期分布图

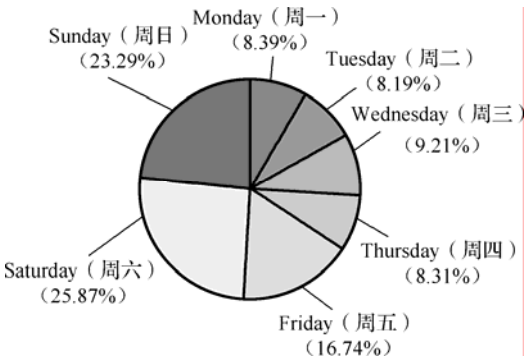


图 3-28　“元信息”更新周周期分布图

以上数据表明发布者主要使用业余时间来发布“元信息”，发布规律符合日常作息周期。根据日周期发布规律，设置网络爬虫的启动时间为 2 点和 18 点，可以及时获取新发布“元信息”。

图 3-29 显示了发布者发布数量分布图，图 3-30 显示了发布者数量与“元信息”数量之间的关系分布图，从图 3-30 可以看出，小部分用户发布了大多数“元信息”，大部分发布者的发布数量保持在较低的水平。发布数量最多的 1.35%用户发布了 68.68%的“元信息”，5.99%的用户发布了 89.65%的元信息，而剩余的 95.01%用户只发布了 10.45%的“元信息”，发布者与发布数量之间符合幂律关系。对于发布数量最多的部分用户来说，每人的发布数量都达到几百甚至几千，属于普通爱好者的概率较低，应该是网站为了保证资源更新速度和服务水平而雇佣的兼职发布人员。所有发布者总数为 741，但是网站用户总数为 261091，发布者只占用户总数的 0.28%，说明只有极少数用户会发布“元信息”，“搭便车”现象非常严重。

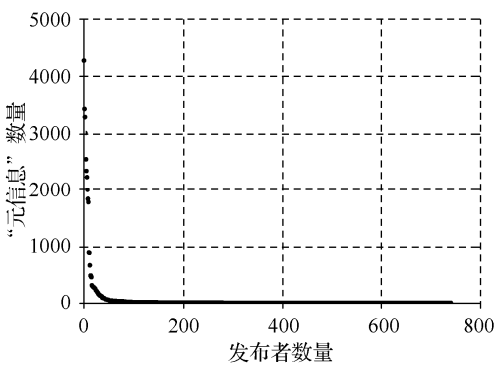


图 3-29　“元信息”发布者发布数量分布图

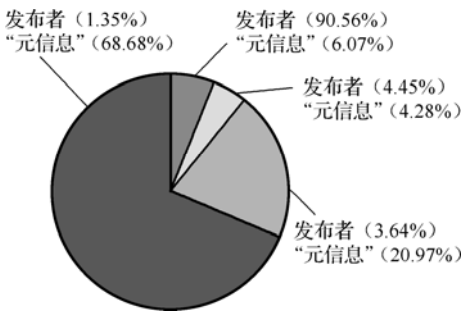


图 3-30　“元信息”发布关系分布图

3.“元信息”流行度

在 SEInR 模型中，感染概率 v 对传播模型的影响非常大，该参数主要与特定信息的被关注程度和流行度有关，特定信息的流行度越高，v 值越大，流行度越低，v 值越小。因此，通过对特定信息流行度的分析，有助于确定 SEInR 模型中的感染概率 v。

下面以“元信息”的回复次数和浏览次数作为流行度分析的主要依据，以发布时间和最后回复时间作为持续时间分析的主要依据。

图 3-31 和图 3-32 分别显示了不同分类的平均回复次数和平均浏览次数，从图 3-31 可以看出，高清影视和普通影视的回复次数远远多于其他分类，说明用户对电影最感兴趣。对于高清影视来说，平均每 40 次浏览就会有一次回复，考虑一个用户对同一页面会浏览多次的情况，高清影视分类的用户回复率还是较高的。

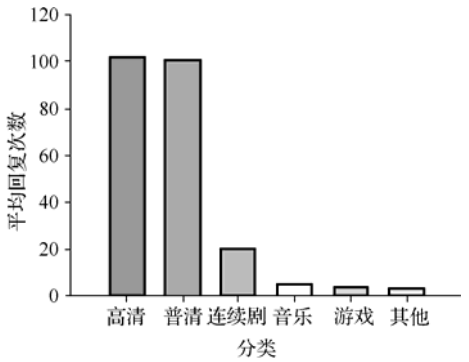


图 3-31　“元信息”平均回复次数比较

“元信息”的持续时长反映了“元信息”流行度。图 3-33 显示了不同分类的平均持续时长，从图 3-33 可以看出，普通影视分类的持续时长最长，“其他”分类被关注程度最低，持续时长最短。图 3-34 显示了普通影视分类持续时长分布图，从图 3-34 可以看出，

该分类的持续时长比较分散，每个不同分段所占数量比较平均。“元信息”流行度与持续时长近似于线性正比关系，如果“元信息”流行度较高，参与用户数量就会较多，浏览次数和回复次数就会不断增长，随着回复次数的增长，持续时长会相应增加。

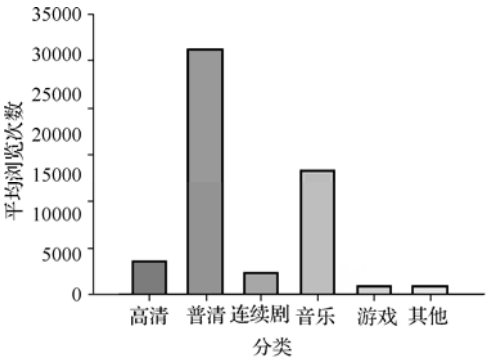


图 3-32　“元信息”平均浏览次数比较

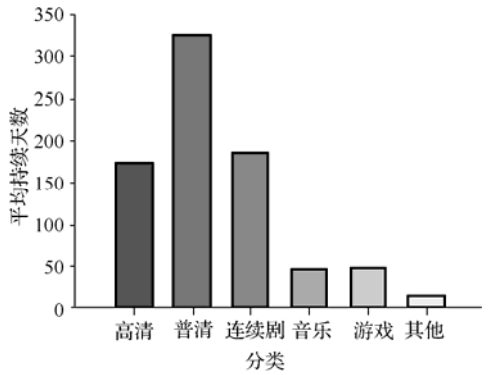


图 3-33　“元信息”不同分类平均持续时长比较

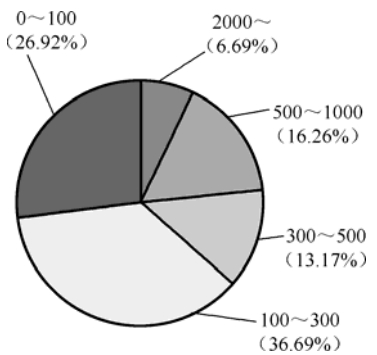


图 3-34　普通影视持续时长分布图

每个网站会根据自身情况和目的采用不同的下载策略和排序策略，这些不同的策略会对“元信息”流行度产生相应影响。如果下载策略使用“回复后下载”，得到的回复次数近似于“元信息”被下载次数，否则，由于回复属于“自愿”行为，得到的回复次数将远远小于被下载次数。网站中“元信息”的排序策略一般分为按发布日期排序和按最后回复日期排序。如果采用按发布日期排序，随着时间流逝，一些流行度较高的“元信息”也会被逐渐遗忘。虽然按照最后回复日期排序可以解决上述问题，始终将流行度较高的“元信息”排在较前位置，但是这种策略会导致网页中的信息不断变化，影响到用户的体验。如何选择适合网站自身的策略，需要在考虑方便性和易用性的基础上，由网站的自身特点和目标群体来决定。

3.4.2　网络拓扑特性分析

复杂网络的拓扑特性对于特定信息传播有着重要影响，相同传播模型在不同拓扑特性上会表现出不同的传播性质。P2P 特定信息传播网络的拓扑特性与传播过程之间相互影响，传播网络是随着特定信息的传播而逐渐形成的一个动态复杂网络。当 P2P 节点开始下载特定信息时，相当于在网络中增加一个节点并建立与现有节点之间的连接关系；当 P2P 节点下载完成并停止共享特定信息时，相当于在网络中删除该节点以及与该节点有关的连接信息。随着传播网络的逐渐增长，网络拓扑特性将会不断发生变化，对特定信息传播产生影响。

为了分析传播网络拓扑特性，需要对互联网中的实际网站进行测量和数据采集。下面以热门视频文件“盗梦空间”为对象进行测量，测量时间从 2010 年 11 月 16 日种子发布时开始，到 2011 年 1 月 25 日结束，持续时间为 70 天，测量间隔时间设置为 10 分钟。通过对已获取的受众信息进行整理、过滤与组织，建立以天为单位的 P2P 特定信息传播网络图，并对传播网络的静态及动态拓扑特性和用户行为进行分析，研究它们对 SEInR 模型的影响。

1．传播过程分析

应用 SEInR 模型对特定信息传播过程进行分析时，主要关注潜伏节点和感染节点的变化过程以及特定信息传播范围。通过潜伏节点和感染节点的分辨方法，对已获取节点进行分类，得到潜伏节点和感染节点的变化数据。特定信息传播范围通过节点累计数量来表示。图 3-35 显示了当前在线潜伏节点和感染节点的变化曲线，从图 3-35 可以看出，传播过程具有明显的分段特性。根据传播特点和外界因素可将传播过程分为如下几个阶段。

（1）瞬时上升阶段。“元信息”发布成功后，根据被关注程度，会有不同数量的节点快速下载“元信息”，并进行文件下载。由于初始传播网络中种子节点数量较少，大部分

节点都处于下载状态，整体下载速度较慢，表现形式为潜伏节点快速增加，感染节点增速较慢。随着传播的进行，部分节点在完成下载后转化为感染节点，整个网络对资源的拥有情况得到较大的改善，整体下载速度加快，潜伏节点转换为感染节点的速度加快，感染节点数量快速上升。

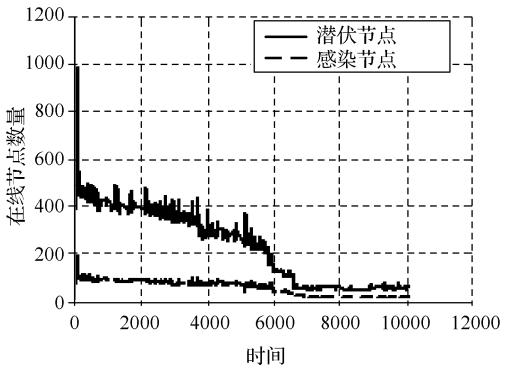


图 3-35　在线节点数量变化图

（2）平稳传播阶段。随着传播的进行，传播网络逐渐显现出以下特点：完成下载的节点在共享特定信息一段时间后，退出传播网络；对该特定信息关注度逐渐降低，新加入的下载节点逐渐减少。此时传播网络中的受众数量保持在一个稳定的水平上，相当于传播模型中的有病平衡点。

（3）逐渐隐退阶段。随着时间的进行，针对该特定信息的“元信息”会逐渐增多，用户选择范围也随之增加，使用被监测“元信息”进行下载的用户数量将逐渐减少。同时，种子节点逐步退出传播网络，网络中的下载节点和上传节点都急速减少，随后稳定在一个较低的水平上。

（4）停止传播阶段。随着种子节点的逐步退出以及参与节点的逐步减少，网络上的资源拥有情况将急剧下降，导致部分文件片在网络上消失，也就是整个网络的资源可用性没有达到 100%，下载节点不能得到完整的特定信息，从而退出下载，整个传播网络处于停止传播阶段。

由于特定信息从开始传播到停止传播所经历的时间很长，特别是从逐渐隐退阶段到停止传播阶段，传播持续时间可长达几年时间。特定信息被关注程度越高，持续时间就越长。由于时间限制，对特定信息“盗梦空间”的监测只关注到逐渐隐退阶段，停止传播阶段的结论是通过对其他被关注度较低的特定信息进行监测和分析得到的。根据 P2P 网络的传播特点，该结论是通用的。

通过对 P2P 特定信息传播过程的分析，结合 SEInR 模型，将整体传播过程使用两阶

段 SEInR 模型来描述：首先根据特定信息被关注程度，设置感染概率和其他参数；当开始进入逐渐隐退阶段时，重新设置模型中的感染概率和其他参数，使得分段后的传播模型能够尽可能地符合实际传播过程。图 3-36 显示了累计节点数量的变化情况，从图 3-36 可以看出，累计数量的变化情况也符合传播过程的阶段性：快速增加、稳定增长和慢速增长。

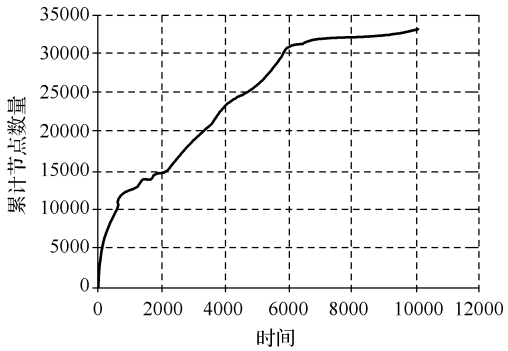


图 3-36　特定信息传播范围变化图

2．节点度值

节点度值是与该节点连接的其他节点数量，它描述了网络局部特性，是刻画节点特性的重要概念。通过对传播网络平均连接度的动态变化情况进行分析，可以了解传播网络的宏观统计特性。复杂网络的平均度 $<k>$ 定义为：

$$<k>=\frac{1}{N_V}\sum_{i=1}^{N_V}\deg(v_i) \tag{3-32}$$

图 3-37 显示了特定信息“盗梦空间”从开始传播到逐渐隐退阶段的平均度变化情况，从图 3-37 可以看出，在平稳传播阶段，网络平均度为 60～80。主要是由于在 P2P 软件中，一般将最大下载节点数和最大上传节点数都设置为 30～50，同时由于节点阻塞协议，使得网络平均度维持在 60～80；在逐渐隐退阶段，由于网络中可供连接的节点数较少，网络平均度为 5～10。

度分布使用节点度的概率分布函数 $P(k)$ 来描述，它表示随机选定一个节点，其度值恰好为 k 的概率，也就是节点有 k 条边的概率。从统计意义上表示为：

$$P(k)=\frac{N_k}{N_V} \tag{3-33}$$

式中，$1\leqslant k\leqslant N_V-1$，$N_k$ 表示网络中度值为 k 的节点数。

度分布函数反映了网络的宏观统计特性。以 2010 年 12 月 1 日的监测数据作为单日分析数据，此时，特定信息已进入平稳传播阶段，选择的数据可代表传播网络的一般规律。

图 3-38 显示了单日传播网络的度分布图，从图 3-38 可以看出，在线节点数量为 753 个，节点的度分布存在明显的长尾特性，度分布部分符合幂律形式 $P(k)\sim Ak^{-\gamma}$，其中，A 为校正系数，为一个常数。度分布的结尾部分与幂律分布偏离较大，主要是由于带宽和软件设置的限制，使得度值不可能无限增长，与理论数据有所偏差。根据文献[6]中介绍的幂律指数估计方法，得到度分布的幂指数 γ 为 2.31，校正系数为 291。根据幂律指数可以得知，传播网络的度分布是不均匀的。对节点度值进一步分析可知：度值最大的一部分节点，大部分属于上传节点，不进行下载，对资源拥有率为 100%，而且长时间在线；而度值处于 40～100 的节点，大部分属于下载节点，下载节点度值处于度值平均值附近。

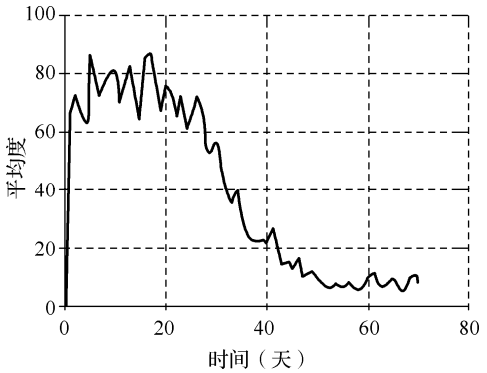


图 3-37　传播网络平均度变化图

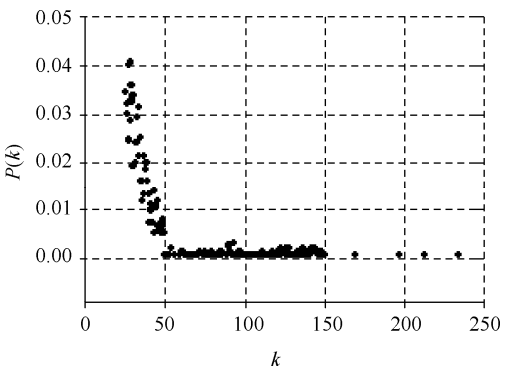


图 3-38　单日传播网络的度分布图

现实世界中的绝大多数复杂网络均表现出异质性，通常描述网络异质性的指标是度分布熵和基尼系数。熵是针对无序性的一种度量，对于给定的一个概率分布 $\{p_1, p_2, \cdots, p_N\}$，相应的熵定义为：

$$E_s = -\sum_{i=1}^{N} p_i \ln(p_i) \tag{3-34}$$

熵刻画了概率分布 $\{p_1, p_2, \cdots, p_N\}$ 的均匀程度，N 个数值 $p_1, p_2, \cdots, p_N$ 相互之间越接近，熵就越大。当 $p_1 = p_2 = \cdots = p_N = 1/N$ 时，熵取得最大值 $\ln N$，而当 $p_1 = 1, p_2 = \cdots = p_N = 0$ 时，熵取得最小值 0。文献[7]定义度分布熵为：

$$E_{\mathrm{Gs}} = -\sum_{k=1}^{N_V - 1} P(k) \ln(P(k)) \tag{3-35}$$

式中，N_V 表示网络中的节点数量。对于每个节点度值均相等的规则网络而言，度分布熵取最小值 0。对于星形网络而言，其度分布熵为：

$$E_{\mathrm{Gs}} = \ln\left(\frac{N_V}{N_V - 1}\right) + \frac{\ln(N_V - 1)}{N_V} \tag{3-36}$$

对于度分布指数为 $\gamma > 0$ 的无标度网络，其度分布熵为[8]：

$$E_{\mathrm{Gs}} = \frac{\gamma \sum_{k=1}^{N_V - 1} (k^{-\gamma} \ln k)}{\sum_{k=1}^{N_V - 1} k^{-\gamma}} + \ln \sum_{k=1}^{N_V - 1} k^{-\gamma} \tag{3-37}$$

洛仑兹曲线是收入不均的一种图形化表示，横坐标为按收入升序排列的累积人口百分比，纵坐标为这些人口收入占总收入的百分比。基尼（Gini）根据该曲线提出了判断收入不均的指标：基尼系数[9]。如图 3-39 所示，基尼系数定义为曲线 OD 与直线 OD 之间的面积(S_A)与三角形 OCD 面积(S_{A+B})的比值，即：$G_n = S_A / S_{A+B}$。

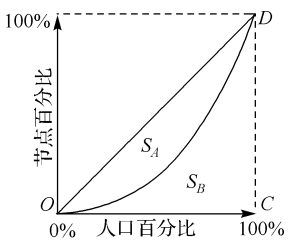


图 3-39　基尼系数定义示意图

将复杂网络的 N_V 个节点按照节点度值从小到大排列，节点 v_i 的度值为 k_i。构造洛仑兹曲线，横坐标为累计节点数与总节点数的比值，纵坐标为累计度值与所有节点总度值的比值，推导出复杂网络的基尼系数表达式为：

$$G_n = \frac{\sum_{i=1}^{[N_V/2]} \left(\left(\frac{N_V - 1}{2} - i \right) (k_{N_V+1-i} - k_i) \right)}{\frac{N_V}{2} \sum_{i=1}^{N_V-1} k_i} \tag{3-38}$$

式中，$[N_V/2]$ 表示不超过 $N_V/2$ 的最大整数。可以发现，节点度值之间的差异越大，基尼系数越大。因此，基尼系数刻画了复杂网络的异质性。对于规则网络，基尼系数为 0；对于星型网络，基尼系数为 $1/2-1/N_V$，由于该网络中只有一个中心节点，而其他节点的度值均相同，因此星形网络的异质性还不够强。

度分布熵是刻画复杂网络中节点度分布的均匀性指标，度值取值范围越广，度值在节点中分布越均匀，则度分布熵越大。基尼系数是刻画复杂网络中各个节点度值的不均匀性指标，它们都可以用来刻画复杂网络的异质性。使用图 3-38 的数据来计算这两个异质性指标参数，其度分布熵为 4.179，基尼系数为 0.715，表明传播网络的异质性较强。

3．路径长度

平均路径长度是度量传播网络特性的重要指标，其定义为网络中任意两个节点之间路径长度 d_{ij} 的平均值，即：

$$< d > = \frac{1}{\frac{1}{2} N_V (N_V - 1)} \sum_{i,j=1, i \neq j}^{N_V} d_{ij} \tag{3-39}$$

平均路径长度表示了传播网络深度，为了加快特定信息传播，对于相同规模的网络，平均路径长度越小越好。图 3-40 显示了每日传播网络平均路径长度变化情况，从图 3-40 可以看出，虽然每日在线节点数量较大，但是网络平均路径长度较短。在平稳传播阶段，平均路径长度始终保持在 3.1～3.5；当传播进入逐渐隐退阶段时，平均路径长度随着网络规模的减少而变小，稳定在 2.7 左右。

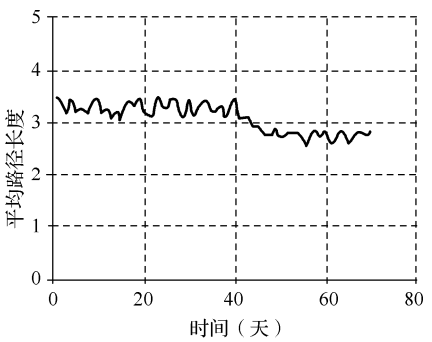


图 3-40　每日传播网络平均路径长度变化图

图 3-41 显示了单日监测数据中路径长度分布情况，从图 3-41 可以看出，节点间路径长度分布近似于符合泊松分布，路径长度主要集中在平均值 3.16 左右。泊松分布左侧部分的数据非常重要，它们表示节点路径长度较小，信息传播速度较快。

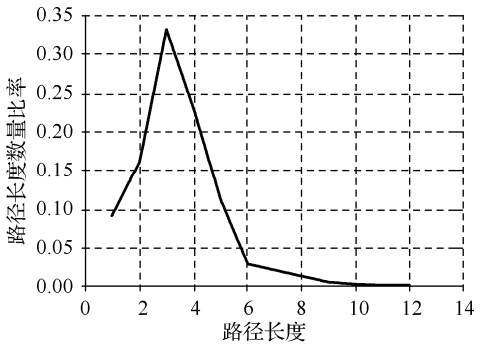


图 3-41　单日监测数据传播网络路径长度分布图

4．节点聚类系数

聚类系数用来衡量一个节点与邻居节点之间的关联程度，在社会学中也称为传递性，是衡量复杂网络中节点之间连接紧密程度的重要指标，它反映了网络中三角形结构的密度。节点 v_i 的聚类系数 c_i 定义为：

$$c_i = \frac{E(\Gamma(v_i))}{\frac{1}{2}k_i(k_i-1)} \tag{3-40}$$

式中，k_i 为节点 v_i 的度，$\Gamma(v_i)$ 为节点 v_i 的邻居节点所形成的子图，$E(\Gamma(v_i))$ 表示 $\Gamma(v_i)$ 中的边数，也就是节点 v_i 的 k_i 个邻居节点之间实际存在的边数。

传播网络的聚类系数定义为所有节点聚类系数的平均值，它代表了网络节点之间连接的紧密程度，数值越大，连接越紧密。图 3-42 显示了传播网络的平均聚类系数变化情况，与其他拓扑特性相同，随着特定信息的传播，网络聚类系数在平稳传播阶段稳定在 0.52 左右；在逐渐隐退阶段，由于种子节点的逐渐退出，传播网络的聚类系数逐渐减小，最终稳定在 0.35 左右。

图 3-43 显示了单日监测数据中每个节点的聚类系数分布情况，从图 3-43 可以看出，大部分节点的聚类系数集中在 0.3～0.7。通过对路径长度和聚类系数的分析可以看到，P2P 特定信息传播网络的平均路径长度在 3.1～3.5，聚类系数在 0.52 左右，相同规模的规则网络的聚类系数为 0.0006 左右。根据小世界网络拓扑特性可知，P2P 特定信息传播网络完全符合小世界网络拓扑特性。

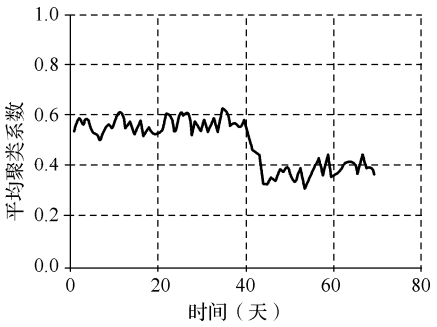


图 3-42　传播网络平均聚类系数变化图

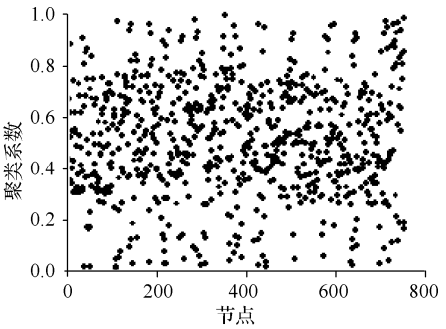


图 3-43　单日监测数据中传播网络节点聚类系数分布图

5．度相关性

度相关性用于考查节点连接时相互选择的偏好性，如果度值大的节点倾向于和度值大的节点连接，则网络是正相关的；反之，网络是负相关的。在拓扑特性分析中，度相关性具有十分重要的意义。按照定义计算度相关性非常复杂，且计算量大。文献[10]给出了一种简便的度相关性计算方法，只需计算节点度值的 Pearson 相关系数 r 即可。

$$r=\frac{\frac{1}{N_E}\sum_{i=1}^{N_E}k_{i1}k_{i2}-\left[\frac{1}{N_E}\sum_{i=1}^{N_E}\frac{1}{2}(k_{i1}+k_{i2})\right]^2}{\frac{1}{N_E}\sum_{i=1}^{N_E}\frac{1}{2}(k_{i1}^2+k_{i2}^2)-\left[\frac{1}{N_E}\sum_{i=1}^{N_E}\frac{1}{2}(k_{i1}+k_{i2})\right]^2} \tag{3-41}$$

式中，N_E 表示复杂网络的总边数，$1\leqslant i\leqslant N_E$，$k_{i1}$ 和 k_{i2} 表示第 i 条边的两个顶点 v_{i1} 和 v_{i2} 的度值。r 的取值范围为 $-1\leqslant r\leqslant 1$，当 $r>0$ 时，网络是正相关的；当 $r<0$ 时，网络是负相关的；当 $r=0$ 时，网络是不相关的。

按照 r 的计算公式，对特定信息每日传播网络进行计算，得到传播网络的度相关系数

变化情况如图 3-44 所示，从图 3-44 可以看出，在传播开始阶段，度相关性 $r>0$，表示此时的传播网络是正相关的，度值大的节点倾向于和度值大的节点连接；随着传播的进行，在平稳传播阶段，度相关性 $r<0$，此时的传播网络已转换为负相关性网络，度值大的节点倾向于和度值小的节点连接。对特定信息传播过程分析可知：在开始传播阶段，特定信息从种子节点向度值大的节点进行传输，为了加快传输速度，此时度值大的节点选择度值大的节点进行连接和传输。在进入平稳传播阶段后，网络中度值大的节点已经完全拥有特定信息，不需要再进行下载，因此其选择度值小的节点来加快整个网络的传输速度。

传播网络的度相关系数与网络的抗毁性有着密切关系。研究表明，对于正相关性网络，高连接度顶点通常群聚在网络的某个局部，将其移走对网络并没有太大影响，网络依然保持较高的连通度。但对于负相关性网络，移走少数高连接度顶点对网络具有毁灭性破坏，因为这些高连接度顶点广泛分布在网络中，与其他顶点间形成很多通路，是网络连通度的重要支撑。

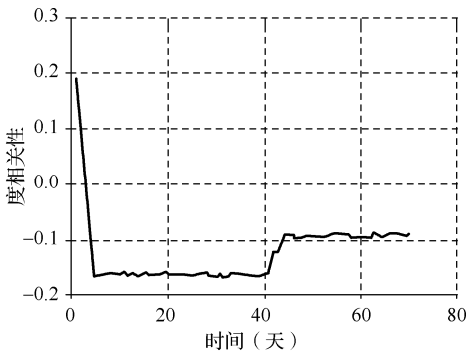


图 3-44　传播网络度相关性变化图

3.4.3　用户行为分析

用户行为是用户在使用 P2P 软件过程中所表现出来的网络行为，它的主要影响因素包括：用户使用习惯、P2P 协议规范以及 P2P 软件实现方式等，表现形式主要有：用户的上下线频率、在线节点的日周期特性、用户在线时长、对特定信息共享时长、节点间会话时长、节点下载速度、节点地址分布以及节点可用性等，这些用户行为会极大地影响 SEInR 模型中相应参数的设置。通过获取的受众信息对用户行为进行分析，有助于理解 P2P 特定信息在传播过程中的内在规律，并对 SEInR 模型的参数设置提供数据参考。

1．在线节点的日周期特性

在对“元信息”发布规律进行分析时可知，“元信息”发布具有明显的周期性，主要

在 18～0 点之间进行发布。在线节点数量和“元信息”发布相类似，也体现了用户使用 P2P 软件的时间习惯。将一天分割为 4 段：0～6 点、6～12 点、12～18 点和 18～0 点，图 3-45 中的(a)和(b)分别显示了在平稳传播阶段和逐渐隐退阶段连续 5 天时间内，各个时间段在线节点数量的变化情况。

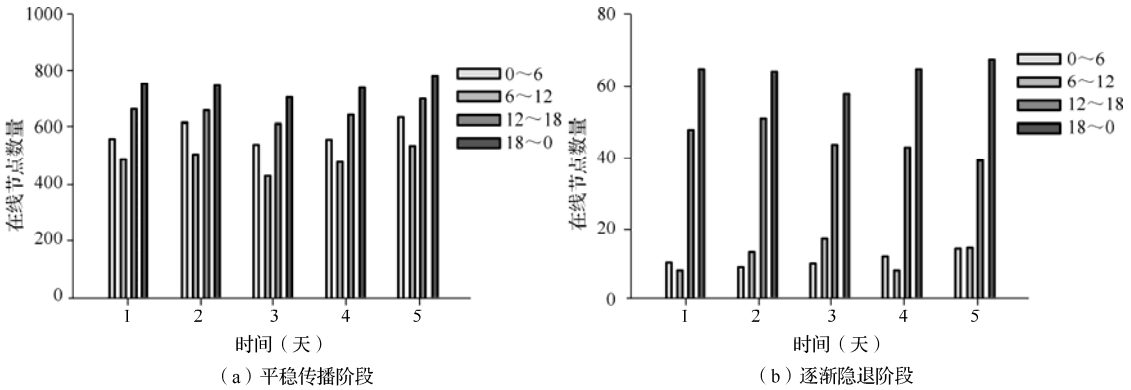


图 3-45　在线节点的日周期特性比较图

从图 3-45 可以看出，在平稳传播阶段，特定信息属于热门资源，下载用户数较多，时间跨度大，具有一定的日周期性；而在逐渐隐退阶段，随着时间推移，特定信息流行度不断下降，下载节点数量逐渐减少，日周期性表现比较明显。因此，在线节点的日周期性不但与用户操作行为有关，而且与特定信息流行度有着密切关系，特定信息流行度越高，日周期性越不明显。在一个热门资源的开始传播阶段，日周期性比较不明显。

2．节点上线频率及在线时长分布

为了统计节点在线时长，需要对受众信息中的节点上下线时间点进行分析。节点上线时间点可以定义为节点首次出现的时间点或者下线后再次出现的时间点。由于主动测量模型每次得到的节点列表不一定是全部节点信息，当节点不出现在某次返回的节点列表中时，不能确定该节点是否已经离线。因此需要进行如下的判断：假设每次返回节点列表数量最大为 C_{Bmax}，对该特定信息进行主动监测得到的在线节点总数量为 C_{Oall}，设定阈值 $\varphi_{\mathrm{Offine}}$ 为：

$$\varphi_{\mathrm{Offine}} = 2\frac{C_{\mathrm{Oall}}}{C_{\mathrm{Bmax}}} \tag{3-42}$$

当某个节点连续 $\varphi_{\mathrm{Offine}}$ 次都没有出现在返回的节点列表中时，可以判断该节点已经离线，离线时间点为该节点最后出现在返回节点列表中的时间点。获取节点状态时，如果节点不可连接，则说明由于 Tracker 服务器没有及时更新缓存数据而将已经离线的节点返回，这时可以直接判断该节点已经离线。

通过对获取的受众信息进行分析，统计节点上线次数，并对节点在整个传输过程的上线次数分布进行研究，判断特定信息传输过程中的节点上线性质。图 3-46 显示了上线次数与节点数之间的对应关系，从图 3-46 可以看出，节点在整个传输过程中的上线次数并不十分频繁，平均上线次数为 3.15，上线次数为 2 的节点数量最多，最大的节点上线次数为 68 次，但是节点数只有 1 个，而且该节点的累积在线时间最长。形成这种现象的主要原因如下。

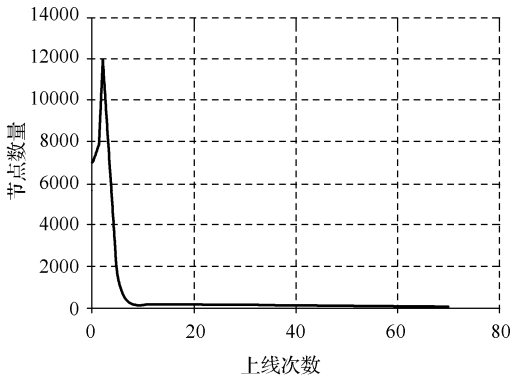


图 3-46　节点上线次数分布图

（1）节点下载特定信息所需的时间较短。特定信息“盗梦空间”平均下载时间为 2 小时，只有 12 次测量机会，而且大部分节点属于“搭便车”节点；

（2）测量时间间隔较长。由于测量间隔时间为 10 分钟，在 10 分钟内的离线及上线在测量中无法显示出来；

（3）测量方式不同。当采用基于 Tracker 服务器日志分析的测量方式，能够获得比较准确的节点上下线时间，而主动测量模型需要频繁地获取节点信息。

在 SEInR 模型中，根据被治愈概率将感染节点分为多个感染子类，每个子类具有不同的治愈概率。治愈概率的设置主要与节点的在线共享时长有关，在线共享时长越长，治愈概率越低，在线共享时长越短，治愈概率越大。因此，节点的在线共享时长对于设置 SEInR 模型中的治愈概率参数具有重要的参考价值。

与分析节点上线次数相类似，需要统计所有节点的在线时长，同时为了得到节点的在线共享时长数据。在统计时，根据节点的资源拥有率判断节点是否处于共享阶段，并对处于共享阶段的在线时长进行单独统计。图 3-47 显示了在线时长、共享时长与节点数量的对应关系，由于时长大于 2 天的节点数量极少，只占总节点数量的 1%左右，为了重点关注大部分数据的规律，图形中只显示了时间单位在 0～300 的数据，从图 3-47 可以看出，大部分节点的在线时长和共享时长较小，共享时长曲线顶点的横坐标为 0，表示这些节点

下载完成后，直接退出传播网络，是典型的“搭便车”行为。对共享时长分析后可知：共享时长为 0 的节点比例为 10.62%，共享时长为 1～6（1 小时内）的节点比例为 29.69%，共享时长为 7～72（6 小时内）的节点比例为 37.66%，共享时长为 73～144 （第 1 天）的节点比例为 14.31%，共享时长为 145～288（第 2 天）的节点比例为 6.26%，共享时长超过 2 天的节点比例仅为 1.46%。对比 SEInR 模型，将感染节点分为 3 个感染子类时，可将共享时长在 1 小时之内的节点集合作为第一个子类，为“搭便车”节点子类；共享时长超过 1 小时但在 1 天之内的节点集合作为第二个子类，为一般共享子类；共享时长超过一天的节点集合作为第三个子类，为长时间共享子类。

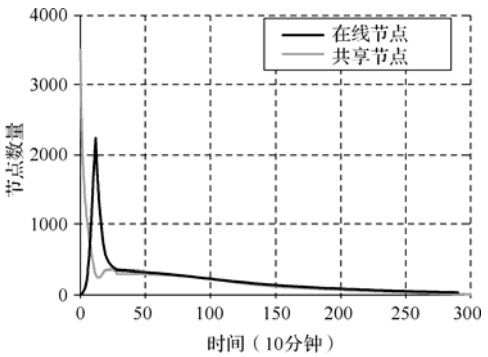


图 3-47　节点在线时长分布图

上线次数多的节点是否具有较长的在线时长，可以通过变量之间的相关性进行分析。设随机变量 X，Y 的样本值分别为 $(x_1,x_2,\cdots,x_n),(y_1,y_2,\cdots,y_n)$，则随机变量 X 和 Y 之间的相关系数 R_{XY} 定义为：

$$R_{XY}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i-\bar{x})^2\cdot\sum_{i=0}^{n}(y_i-\bar{y})^2}} \tag{3-43}$$

式中，n 表示样本值个数。R_{XY} 的取值范围为 $-1\leqslant R_{XY}\leqslant 1$，其性质如下：当 $R_{XY}>0$ 时，两个变量正相关；$R_{XY}<0$ 时，两个变量负相关；$R_{XY}=1$ 时，2 个变量完全线性相关；$R_{XY}=0$ 时，2 个变量无线性相关关系。当 $0<|R_{XY}|<1$ 时，表示 2 个变量存在一定程度的线性相关，且 $|R_{XY}|$ 越接近 1，2 个变量间的线性关系越密切，$|R_{XY}|$ 越接近 0，2 个变量间的线性关系越弱，一般可按照 3 级划分：$|R_{XY}|<0.4$ 为低度线性相关，$0.4\leqslant|R_{XY}|<0.7$ 为显著线性相关，$0.7\leqslant|R_{XY}|<1$ 为高度线性相关。对节点的上线次数和在线时长进行相关系数计算，可以得到 $R_{XY}=0.31$，上线次数与在线时长之间存在着一定的线性关系，但相关性不是很密切。

3．节点间会话时长分布

P2P 文件共享系统中，每个节点在下载文件的过程中，会根据 P2P 网络的返回节点列表与多个节点之间建立连接并进行数据传输。下载节点并不是与节点列表中所有节点都建立连接关系，而是根据节点选择策略，选择合适的节点建立连接。在传输过程中，根据节点阻塞算法对已经建立的连接进行判断，决定是否需要断开连接和阻塞对方节点。节点间会话时长表示连接的持续时长。由于阻塞算法的作用，已经建立的连接会被中断，导致节点间会话时长的随机性。图 3-48 显示了单日监测数据中的会话时长分布情况，从图 3-48 可以看出，会话时长为 1 时，连接所占比例较高，但是会话时长为 2 时，连接所占比例比较低，然后开始呈现逐渐上升再逐渐下降的过程，在会话时长为 13 时，连接所占比例最高。

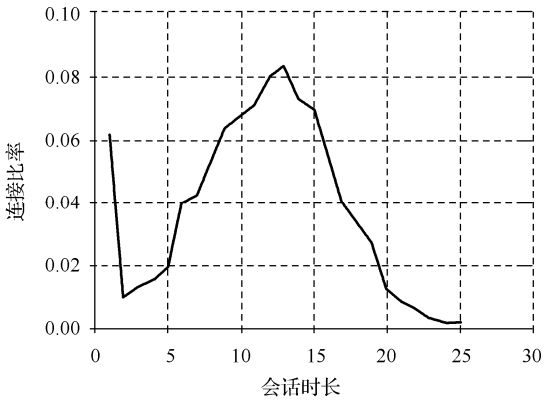


图 3-48　节点间会话时长分布图

对特定信息的传播过程和阻塞算法分析可知，在节点间连接建立后，对于传输速度较慢的连接，阻塞算法会很快将其断开，以使下载节点可以选择其他具有更快传输速度的节点进行连接，会话时长为 1 的连接大部分属于这种情况。当连接没有很快被阻塞时，表示该连接已经进入正常传输阶段，会话时长越长，被阻塞的概率越低。需要说明的是，图 3-48 显示的是热门资源的情况，对于受众信息较少的冷门资源，由于节点间的传输速度较慢，会出现连接被频繁阻塞的情况，会话时长较短的连接所占比例较大。

4．节点下载速度分布

在 SEInR 模型中，节点下载速度对模型中的参数 p_{ER} 和 p_{EI} 有着重要影响，下载速度越快，潜伏节点转换为感染节点的速度越快，p_{ER} 越小，p_{EI} 越大。因此，对传播网络的下载速度进行分析，可以为设置 p_{ER} 和 p_{EI} 提供数据参考。

主动测量模型只能得到传播网络拓扑结构，难以得到节点下载速度，而被动测量模型由于测量范围的限制，难以得到被监测网络外的节点下载速度。通过对 P2P 协议的分

析，可以采用一种间接方法：使用特定信息文件大小和节点的资源拥有率变化情况，间接计算出节点平均下载速度。虽然得到的平均下载速度是一个统计分析值，但是对于分析P2P 特定信息传播规律有一定的作用。具体计算方法为：从“元信息”中得到特定信息的大小为 S_o M，对指定节点进行 n 次状态测量，该节点的资源拥有率变化情况为 $p_{o1},p_{o2},\cdots,p_{on}$，并且 $0\leqslant p_{o1}\leqslant p_{o2}\leqslant\cdots\leqslant p_{on}\leqslant 1$，测量间隔时间为 T_o 分钟，则该节点在任意两次测量时间点 $(p_{\mathrm{oi}},p_{\mathrm{oj}})$ 之间的平均下载速度为：

$$D_{\mathrm{SAij}}=\frac{(P_{\mathrm{oj}}-P_{\mathrm{oi}})S_o\cdot 1024}{(j-i)T_o\cdot 60}\mathrm{KB/s} \tag{3-44}$$

图 3-49 显示了单日监测数据中节点平均下载速度分布情况，从图 3-49 可以看出，当下载速度在 190KB/s 左右时，节点所占比例较高，大部分节点的下载速度集中在 100～450KB/s。当下载速度低于 50KB/s 或者高于 450KB/s 时，节点所占比例明显减小，说明在现阶段使用低速网络上网的用户比例已经非常小，但是使用光纤等高速网络线路上网的用户比例也比较小。所有节点的平均下载速度为 286KB/s，相当于 2.2Mb/s 带宽的下载速度，这主要应归结于电信服务商将网络带宽从原来的 2Mb/s 提高到 4Mb/s。从中可以看出，使用 4Mb/s 带宽的用户比例并不是很高，还有待于进一步发展。根据所获得的平均下载速度和特定信息文件大小，可以为设置 SEInR 模型中的参数 p_{ER} 与 p_{EI} 提供参考依据。

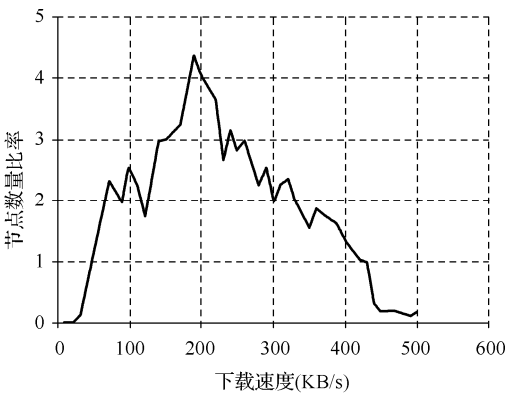


图 3-49　节点平均下载速度分布图

5．节点地址分布

节点位置信息是受众信息中最重要的信息，位置信息包含了 IP 地址和端口号，对 IP 地址进行分析，可以知道参与特定信息传播的节点在位置上的分布情况。下面使用 IP 地址的随机测度对 IP 地址分布进行分析。

熵刻画了概率分布的均匀程度，对 IP 地址熵的定义为：假设 IP 地址集合中有 n 个元素，它们分属于 m 个子网，子网的定义为前 16 位相同的 IP 地址集合。第 i 个子网在 IP 地址集合中出现的概率为 $p_i = n_i / n$，$i = 1,2,\cdots,m$。其中，n_i 表示第 i 个子网中 IP 地址的数量，$p_i \geqslant 0$，且 $p_1 + p_2 + \cdots + p_m = 1$。则 IP 地址集合的 IP 地址熵定义为：

$$E_{s\mathrm{IP}} = -\sum_{i=1}^{m} p_i \ln(p_i) \tag{3-45}$$

如果 $m = 1$，所有 IP 地址在一个子网内，IP 地址熵 $E_{s\mathrm{IP}} = 0$，取得最小值。如果每个 IP 地址都以相同概率出现，即 $m = n$，$p_1 = p_2 = \cdots = p_n = 1/n$，每个 IP 地址就是一个子网，IP 地址熵 $E_{s\mathrm{IPmax}} = \ln n$，取得最大值。IP 地址的随机测度 E_{IP} 定义为 IP 地址熵与最大 IP 地址熵的比值，即：

$$E_{\mathrm{IP}} = \frac{E_{s\mathrm{IP}}}{E_{s\mathrm{IPmax}}} \tag{3-46}$$

由 E_{IP} 定义可知，$0 \leqslant E_{\mathrm{IP}} \leqslant 1$，表示 IP 地址的随机程度。$E_{\mathrm{IP}}$ 越接近 1，IP 地址的随机性越大；E_{IP} 越接近 0，IP 地址的随机性越小。图 3-50 显示了每日传播网络的 IP 地址随机测度变化情况，从图 3-50 可以看出，IP 地址的随机测度始终保持在较高水平，表示受众信息中的节点 IP 地址比较分散、随机程度高。

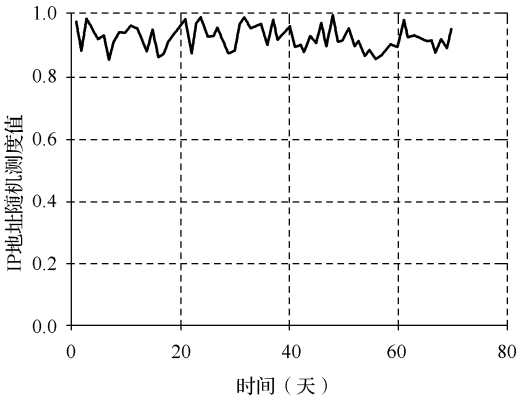


图 3-50　IP 地址随机测度变化图

P2P 下载的一个重要特点是采用了分片机制，将原始文件分成不同文件块和文件片进行传输，使得特定信息能够快速传播到多个节点中，使原来的串行下载变成了并行下载。从连接角度看，使少量连接变成了多条连接。P2P 网络是一个由很多节点组成的网络，每个节点的来源位置比较随机，没有任何规律性。这些特点导致了 IP 地址的随机测度比较大。

6．节点可用性

节点可用性是衡量网络系统性能的重要指标，这里的节点可用性是指整体网络的节点可用性。在每次获得返回节点列表后，与每个节点进行连接，获取该节点状态信息。如果收到节点的状态返回信息，表示节点可连接，否则，表示节点不可连接。整体网络的节点可用性U_R定义为：

$$U_R = \frac{N_{\mathrm{CR}}}{N_{\mathrm{AF}}} \tag{3-47}$$

式中，N_{CR}为一段时间内所有可连接节点数量，N_{AF}为相同时间内总节点数。图 3-51 显示了每日传播网络的整体节点可用性变化情况，从图 3-51 可以看出，在传播开始阶段，由于 P2P 节点的持续参与程度较高，节点状态比较稳定，节点可用性较高，随着时间的推移，传播节点稳定性不断下降，节点可用性逐渐降低，并稳定在较低的数值上。

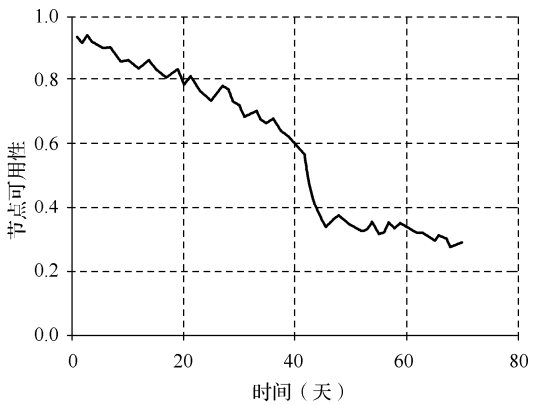


图 3-51　传播网络节点可用性变化图

综上所述，通过对“元信息”属性、网络拓扑特性以及用户行为的分析，可以为 SEInR 模型参数设置提供数据支持。

从“元信息”属性分析结果可知，不同的文件分类具有不同的文件大小分布；视频类“元信息”发布最为活跃，持续时间较长，特别是高清影视，用户参与度较高。通过“元信息”属性分析，可以为 SEInR 模型中的感染概率设置提供参考依据。

从网络拓扑特性分析结果可知，特定信息传播过程一般分为 4 个阶段：瞬时上升阶段、平稳传播阶段、逐渐隐退阶段和停止传播阶段；传播网络具有典型的小世界特性；节点度分布存在明显的长尾特性，部分符合幂律形式，异质性较强；稳定阶段的传播网络为负相关网络。

从用户行为结果分析可知，不同传播阶段的在线节点具有不同的日周期特性；节点上线次数与在线时长之间存在着一定的线性关系，但相关性不是很密切；由于 P2P 协议中

阻塞算法的作用，会话时长为 1 的连接数量较多；IP 地址的随机测度始终保持在较高水平，表示 IP 地址比较分散、随机程度高；整个网络的节点可用性随着时间推移，逐渐下降，并且在开始传播的瞬时上升阶段，整个网络的节点可用性最高。

3.5　P2P 特定信息传播控制

随着 P2P 文件共享系统的发展，在 P2P 网络中存在着大量的侵权、非法、色情等不良信息。这些不良信息的传播不仅影响到社会和谐稳定、网络文化安全以及青少年身心健康，还会占用大量的网络带宽，加剧网络拥堵。

目前，P2P 网络信息传播控制主要采用整体控制技术和文件污染技术，整体控制技术由网络服务运营商通过 P2P 流量识别和封堵对特定的 P2P 网络实施整体控制，而不是针对 P2P 特定信息传播进行控制，容易引起广大 P2P 用户的不满。文件污染技术通过对特定文件进行污染，干扰文件下载过程，降低文件下载成功率，影响用户体验。这两种控制方法都属于非精细化控制技术，难以实现对 P2P 特定信息传播进行有效控制。

下面给出了一个 P2P 特定信息传播控制模型，其基本思路是：首先通过“元信息”分类来辨别需要监控的不良信息，并通过所获取的受众信息判断传播网络状态，选择需要控制的传播网络；然后根据传播网络控制策略选择需要控制的目标节点；最后根据 P2P 节点控制方法对目标节点实施控制。

3.5.1　传播控制模型框架

该模型由“元信息”分类、传播网络状态判断、控制目标节点选择以及节点控制等部分组成，模型框架如图 3-52 所示。

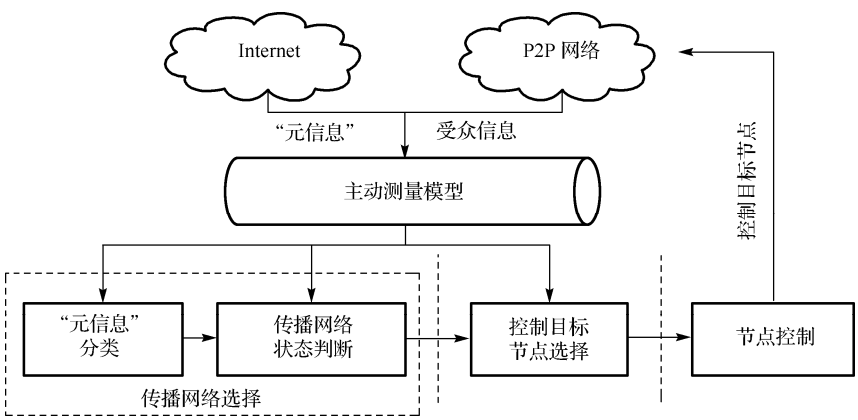


图 3-52　P2P 特定信息传播控制模型框架

模型的主要模块功能描述如下：

（1）“元信息”分类。由于网络爬虫技术和网页信息限制，获取的“元信息”中既包括不良信息，也包括正常信息。为了准确、高效地对“元信息”进行分类，采用改进的支持向量机（SVM）算法对“元信息”进行分类，解决“元信息”分类时关键词特征稀疏和样本高度不均衡问题；

（2）网络状态判断。根据“元信息”分类结果，采用主动测量模型对不良信息进行监测，获得不良信息传播网络变化情况。结合 SEInR 模型和 P2P 特定信息传播网络拓扑特性及用户行为，对不良信息传播网络状态进行分析和判断，得到正在 P2P 网络上传播的不良信息列表，然后选择需要控制的 P2P 特定信息传播网络；

（3）目标节点选择。研究适合于 P2P 特定信息传播网络的控制策略，针对需要控制的特定信息传播网络，根据控制策略选择控制目标节点；

（4）节点控制。针对获得的控制目标节点，使用基于 P2P 协议的节点控制方法对其实施控制操作，破坏特定信息传播网络的鲁棒性，抑制特定信息的传播。

“元信息”分类属于文本分类识别问题，国内外开展了大量的研究，这里不再赘述。网络状态判断中所涉及的模型和方法前面已经做了详细介绍。因此，下面重点讨论目标节点选择策略和节点控制方法。

3.5.2 目标节点选择策略

1．传播网络鲁棒性

鲁棒性是指一个系统的内部结构或外部环境发生改变时，能够维持其功能的能力。P2P 特定信息传播网络的鲁棒性是指当传播网络中部分节点或连接被破坏时，特定信息仍然能够继续维持传播的能力。最早的鲁棒性研究是为了防止疾病传播，例如，在网络中“删除”节点相当于个体为预防疾病而接种，由于接种不仅可以阻止被接种个体感染疾病，同时还破坏了个体之间疾病传播的路径，相当于在网络中某些节点出现故障后，其鲁棒性高低的问题。

传播网络鲁棒性包括两个方面：对随机破坏的容错能力和对蓄意攻击的抗毁能力。从网络在受到攻击时流量是否发生转移来看，可以将鲁棒性分为静态鲁棒性和动态鲁棒性。静态鲁棒性是指当删除网络中的节点时，不需要重新分配网络上的流量，网络能够保持其功能的能力。动态鲁棒性是指当删除网络中的节点时，网络上的流量需要重新分配，经过动态平衡后，网络仍能维持其功能的能力。在动态鲁棒性中，如果少数节点或边发生故障，这种故障通过节点之间的耦合关系引起其他节点发生故障，产生连锁效应，最终导致相当一部分节点甚至整个网络崩溃，这种现象称为级联失效。文献[11]研究了随机破坏与

蓄意攻击对随机网络和无标度网络的影响，对于随机网络来说，使用随机破坏和蓄意攻击的效果几乎一样，随着失效节点的增加，随机网络的平均最短路径略有上升；而对于无标度网络来说，由于网络度值分布的极端非均匀性，使用随机破坏时，平均最短路径几乎没有什么变化。但是当蓄意攻击网络中度值很大的节点时，平均最短路径陡然上升，表现出较强的抗随机破坏能力和较差的抗蓄意攻击能力。文献[12]通过对无标度网络下的相变性质的研究，得到无标度网络存在最大连通集团的判定标准为：$<k^2>/<k>=2$。当以概率 $p(0\leqslant p\leqslant 1)$ 随机删除部分节点及其边时，度分布 $P(k)$ 会转化为新的度分布 $\tilde{P}(k)$，即：

$$\tilde{P}(k)=\sum_{k_c\geqslant k}^{k_{\max}} P(kc)\begin{bmatrix}k_c\\k\end{bmatrix}(1-P)^k P^{(k_c-k)} \tag{3-48}$$

因此，概率 p 的阈值 p_c 可以表达为：

$$p_c=1-\frac{1}{\dfrac{<k_c^2>}{<k_c>}-1} \tag{3-49}$$

当 $p>p_c$ 时，无标度网络的鲁棒性被破坏。由于 $<k_c^2>/<k_c>$ 数值较大，可知对于无标度网络，随机破坏的概率阈值 p_c 较大，只有删除大部分节点和连接时，才可以破坏无标度网络的鲁棒性。该网络对随机破坏具有较强的抗攻击能力。

蓄意攻击将依次删除网络中度值最大节点及其边，最大度 $k_{\max}$ 可以估计为：

$$\sum_{k=k_{\max}}^{\infty} P(k)=\frac{1}{N_V} \tag{3-50}$$

蓄意攻击后，网络新的最大度 $\tilde{k}_{\max}$ 可以估计为：

$$\sum_{k=\tilde{k}_{\max}}^{k_{\max}} P(k)=\sum_{k=\tilde{k}_{\max}}^{\infty} P(k)-\frac{1}{N_V}=p \tag{3-51}$$

由于幂律分布可以表示为 $P(k)=Ak^{-\gamma}$，参数 A 的近似计算公式为：

$$A=(k_{\min})^{\gamma-1}(\gamma-1) \tag{3-52}$$

式中，$k_{\min}$ 为网络中最小度。因此，可知蓄意攻击的概率阈值 p_c 为：

$$p_c=\left(\frac{\tilde{k}_{\max}}{k_{\min}}\right)^{1-\gamma}-\frac{1}{N_V} \tag{3-53}$$

通过对公式（3-53）的分析可知，蓄意攻击的概率阈值 p_c 较小。因此，只要删除极少数度值最大节点即可破坏无标度网络的鲁棒性。

2．基本免疫策略

免疫策略是传播动力学中的重要概念，通过应用不同的免疫策略，可以对传染病传播产生影响，抑制传染病的传播。在 P2P 文件共享系统中，P2P 特定信息的大范围传播行为与传染病的传播过程相类似，可以参考免疫策略思想来研究 P2P 特定信息传播网络的控制策略。主要的免疫策略有如下几种。

（1）随机免疫：随机免疫是指从网络中随机地选取部分节点进行免疫，平等对待不同度值的节点。根据网络鲁棒性的研究，对于无标度网络，其免疫临界值为 $p_c = 1-1/(<k_c^2>/<k_c>-1)$。随着网络规模增长，$p_c \to 1$，需要对网络中几乎所有节点都实施免疫才能保证消灭传染病传播，其免疫措施不具备高效性；

（2）目标免疫：目标免疫是指根据网络节点度分布的不均匀性，依次选择度值最大的节点进行免疫，一旦这些节点被免疫后，与它们所连的边也可以从网络中去除，使得传播路径大大减少，从而高效地消除传染病扩散。但是这种策略需要事先了解网络全局信息，至少需要对网络中的节点度值有比较清楚的认识；

（3）偏好目标免疫：通过定义指标 α 来刻画免疫对象选择策略，感染节点被治愈的概率和 k^{α} 成正比。正常数 α 越大，对高度值节点的选择偏好就越大。$\alpha = 0$ 时，退化为随机免疫；$\alpha = \infty$ 时，对应于目标免疫大于某一给定度值的所有节点；

（4）熟人免疫：熟人免疫是指从 N_V 个节点中随机选出比例为 p 的节点，再随机选择每个被选节点的直接邻居节点进行免疫，它的出发点是为了回避目标免疫中需要知道全局信息的问题。但是，这种对所有邻居节点不加甄别、随机选取的方式，以及仅限于选取直接邻居的做法，使得熟人免疫的实际效果受到了制约；

（5）基于图覆盖问题的免疫：基于图覆盖问题的免疫是将分散式免疫过程视为一个 d 跳范围内的图覆盖问题，即对于一个网络节点 v_i，寻找与节点 v_i 的距离在 d 跳范围内的具有最高连接度的节点，对其实施免疫。这种免疫方法使用了一定范围内的局部拓扑知识，但是，它没有考虑任何与领域相关的启发知识，且没有回答什么是合适的 d 值以及该如何确定之，因此存在一定的局限性。

3．面向 SEInR 模型的控制策略

在传播动力学理论中，节点免疫是预防传染病传播的重要手段，不同免疫策略有着不同的控制代价和效果。传统免疫思想是对需要免疫的个体提前注射疫苗，使其不被传染病感染。由于 P2P 特定信息传播网络的形成特点，要想提前知道哪些节点参与特定信息传播是困难的，因此只能参考免疫策略思想，在传播网络形成后，选择目标节点进行控制操作，这种方式也称为事后免疫。现有免疫策略中使用较为普遍的是随机免疫和目

标免疫，对应于随机控制策略与目标控制策略。下面从理论角度对随机控制策略和目标控制策略进行分析。

1）随机控制策略分析

随机控制策略是指随机地抽取 P2P 特定信息传播网络中的节点及其连接进行控制，假定随机抽取概率为 p_{m}，结合 SEInR 模型，控制后的模型方程组为：

$$\begin{cases} \dfrac{\mathrm{d}S_R}{\mathrm{d}t} = p_F - vc_u S_R\left(E_R + \sum\limits_{i=1}^{n_I} I_{\mathrm{R}i}\right) - p_F S_R - p_m S_R \\ \dfrac{\mathrm{d}E_R}{\mathrm{d}t} = vc_u S_R\left(E_R + \sum\limits_{i=1}^{n_I} I_{\mathrm{R}i}\right) - (p_{\mathrm{EI}} + p_{\mathrm{ER}} + p_F)E_R - p_m E_R \\ \dfrac{\mathrm{d}I_{\mathrm{R}i}}{\mathrm{d}t} = p_{\mathrm{EI}} p_{\mathrm{I}i} E_R - (\delta_i + p_F) I_{\mathrm{R}i} - p_m I_{\mathrm{R}i} \\ \dfrac{\mathrm{d}R_R}{\mathrm{d}t} = p_{\mathrm{ER}} E_R + \sum\limits_{i=1}^{n_I} \delta_i I_{\mathrm{R}i} - p_F R_R - p_m R_R \\ I_R(0) = I_0 \end{cases} \tag{3-54}$$

从该模型中可以看出，增加控制策略后的动力学模型相当于提高原有模型的退出概率 p_F。根据基本再生数的计算方法，控制后的传播模型基本再生数为：

$$\tilde{R}_0 = \frac{vc_u}{(P_{\mathrm{EI}} + P_{\mathrm{ER}} + P_F + P_m)}\left(1 + \sum_{i=1}^{n_I} \frac{P_{\mathrm{EI}} P_{\mathrm{I}i}}{(\delta_i + P_F + P_m)}\right) \tag{3-55}$$

从基本再生数 R_0 角度来说，对正在传播的特定信息实施随机控制策略，相当于在原有基本再生数 $R_0 > 1$ 的基础上，通过抽取概率 p_m 的变化，使得控制后的基本再生数 $\tilde{R}_0 < 1$。在从 $R_0 > 1$ 向 $\tilde{R}_0 < 1$ 的转变过程中，只有参数 p_m 可以进行变化，由于 p_m 的取值范围为 $0 \leqslant p_m \leqslant 1$。因此，为了达到阻止特定信息传播的目的，$p_m$ 的值需要尽可能地大，而且原始基本再生数 R_0 越大，p_m 越大。

2）目标控制策略分析

一般的目标控制策略是对传播网络中度值最大的节点进行控制，根据对 SEInR 模型的分析，对特定信息传播起关键作用的是那些长期在线的种子节点，由于这些节点拥有完整的文件信息，并长时间在线对文件进行共享，任何下载节点只要能够连接到这些种子节点就可以完成特定信息下载，达到特定信息传播的目的。根据对受众信息的分析，节点在线时长与度值之间存在着一定的正比例关系。因此对模型中治愈率最低的感染节点子类（种子节点子类）进行目标控制，相应的控制后模型方程组为：

$$\begin{cases} \dfrac{\mathrm{d}S_R}{\mathrm{d}t} = p_F - vc_u S_R\left(E_R + \sum_{i=1}^{n_I} I_{\mathrm{R}i}\right) - p_F S_R \\ \dfrac{\mathrm{d}E_R}{\mathrm{d}t} = vc_u S_R\left(E_R + \sum_{i=1}^{n_I} I_{\mathrm{R}i}\right) - (p_{\mathrm{EI}} + p_{\mathrm{ER}} + p_F)E_R \\ \dfrac{\mathrm{d}I_{\mathrm{R}i}}{\mathrm{d}t} = p_{\mathrm{EI}} p_{\mathrm{I}i} E_R - (\delta_{\mathrm{m}i} + p_F) I_{\mathrm{R}i} \\ \dfrac{\mathrm{d}R_R}{\mathrm{d}t} = p_{\mathrm{ER}} E_R + \sum_{i=1}^{n_I} \delta_i I_{\mathrm{R}i} - p_F R_R \\ I_R(0) = I_0 \end{cases} \tag{3-56}$$

式中，$\delta_{\mathrm{m}i}$ 为控制后的各个感染者子类治愈率，由于种子节点子类被控制，相应的感染者子类治愈率由一个趋近于 0 的值变化为趋近于 1 的值。由于种子节点的感染者子类治愈率趋近于 0，相应的基本再生数 R_0 数值较大，能够保证 $R_0 > 1$，使得特定信息可以继续传播，但是当该子类被控制后，治愈率趋近于 1，使得 R_0 的数值急剧减小，可以有效抑制特定信息的传播。

4．目标节点标识与选择算法

1）目标节点标识方法

通过对上述的控制策略分析可知，随机控制策略对 P2P 特定信息的传播抑制作用较差，而目标控制策略的抑制作用比较好。只要对 P2P 特定信息传播网络中的关键节点进行控制操作，就能有效地抑制特定信息的传播。因此，正确的控制目标应该是传播网络中那些对特定信息传播起着关键作用的节点。在传统的目标控制策略中，关键节点就是度值最大的小部分节点。但是，在 P2P 特定信息传播网络中，由于传播网络的动态性和特殊性，如果仅仅以度值最大为标准来选择目标节点，并不符合特定信息的传播特点，对传播网络的抑制作用比较有限。为了更加准确地选取目标节点，根据 P2P 特定信息传播规律以及传播网络的拓扑特性和用户行为，采用如下的目标节点标识方法：

（1）在线时长较长。在 SEInR 模型中，将感染者分为不同的感染者子类，不同子类具有不同的治愈率。长时间在线节点称为种子节点，这些节点对特定信息传播起着关键作用，由于长时间在线，节点治愈率很低，能够感染更多的节点；

（2）度值较大。通过传播网络的拓扑特性分析可知，节点度值分布符合一定的幂律特征。由于连接度值分布的极端非均匀性，只要有意识地控制少量高度值节点，就会对特定信息的传播产生影响，传播网络直径将增大，传播性能下降；

（3）可用性较强。P2P 网络中的节点不仅在连接度分布上存在极端不均匀性，而且在节点的能力水平、存储性能等方面也存在高度异构性。网络中可能存在这样一部分节点，

尽管其连接度并不高，但是在维持网络功能、保证网络性能等方面是高度可用的。例如，在特定信息传播网络中，那些稳定在线的、具有较高能力水平且能持续返回高质量结果的邻近节点，其可用性和重要性显然要高于普通节点；

（4）拥有完整资源。在特定信息传播过程中，P2P 软件（如 BitTorrent）使用文件片选择算法优先传输最稀缺文件片，使得文件片在网络上的分布比较平均。因此，拥有完整资源的在线节点应当优先控制；

（5）拥有稀缺文件片。文件片是 P2P 网络中最小的传输单位，当一个节点被控制后，与该节点相连接的节点将通过其他节点获取剩余文件片。为了彻底阻止特定信息传播，只需要使网络中存在的文件片不完整即可。因此，拥有稀缺文件片的在线节点是目标节点的重要组成部分。

2）目标节点选择算法

目标节点选择算法将根据目标节点标识方法来选择目标节点，算法的具体实现步骤描述如下：

（1）定义算法的各个参数，主要有在线时长最长的节点比率阈值 p_{yt}、度值最大的节点比率阈值 p_{yk}、节点可用性过滤比率阈值 p_{yu}、目标节点比率阈值 p_{ys} 以及选择间隔时间 T_{sit}；

（2）对当前传播网络中的节点信息按照在线时长进行排序，并根据阈值 p_{yt} 将在线时长最长的部分节点存入集合 U_{yt} 中；

（3）对当前传播网络中的节点信息按照度值大小进行排序，并根据阈值 p_{yk} 将度值最大的部分节点存入集合 U_{yk} 中；

（4）获取当前传播网络中拥有完整资源的节点，并存入集合 U_{yo} 中；

（5）合并集合 U_{yt}、U_{yk} 与 U_{yo} 中的节点数据，形成集合 $U_{\mathrm{ytk}}=U_{\mathrm{yt}}\cup U_{\mathrm{yk}}\cup U_{\mathrm{yo}}$；

（6）对集合 U_{ytk} 中的节点按照可用性进行排序，并将可用性低于 50%的节点中可用性最低部分节点过滤，过滤比率为 p_{yu}；

（7）对所有节点的资源拥有情况进行统计，得到传播网络中数量最少的文件片编号，将拥有该文件片的节点存入集合 U_{ys}，并计算 $U_{\mathrm{ys}}=U_{\mathrm{ys}}-U_{\mathrm{ytk}}$；

（8）根据阈值 p_{ys} 将集合 U_{ys} 中的部分节点加入集合 U_{ytk} 中，使得集合 U_{ytk} 中的节点数量不超过 $p_{\mathrm{ys}}N_V$，N_V 为传播网络中的节点数量；

（9）将 U_{ytk} 中的节点存入目标节点集合 U_{ytka} 中，即 $U_{\mathrm{ytka}}=U_{\mathrm{ytka}}+U_{\mathrm{ytk}}$。由于 P2P 传播网络的动态性，根据间隔时间 T_{sit}，转到步骤（2），重新统计目标节点 U_{ytk}，并将集合 U_{ytk} 中的节点存入 U_{ytka} 中。

下面对算法的时间复杂度进行分析。

该算法中使用最多的是排序算法，线性排序算法的时间复杂度一般为 $O(n^2)$，n 为待排序元素数。步骤（2）和步骤（3）需要对所有节点进行排序，步骤（5）是对集合 U_{ytk} 中的节点进行排序，集合 U_{ytk} 中的节点数量最多为 $(p_{\text{yt}}+p_{\text{yk}})N_{\text{V}}$，步骤（6）的算法复杂度为 $O(N_V s)$，s 为特定信息的文件片数量。因此，该算法的时间复杂度为 $O((2+(p_{\text{yt}}+p_{\text{yk}})^2)N_V^2+N_V s)$。该算法在统计过程中只需比较操作和保存操作，计算量很小，可在短时间内完成。

3.5.3 P2P 节点控制方法

为了控制特定信息的传播，需要对已选择目标节点实施控制操作，将其移出特定信息传播网络，不再参与特定信息的上传和下载。

P2P 节点控制方法的基本思路如下：

（1）使目标节点退出传播网络。模拟目标节点向 Tracker 服务器发送退出传播网络命令。在 BitTorrent 协议中，将参数 event 设置为 stop，向 Tracker 服务器发送 Get 命令即可。为了防止目标节点向 Tracker 服务器发送状态报告消息，当目标节点在可控制网络内部时，可根据关键字和特定信息 Hash 值直接将此消息过滤。否则，当主动测量模型获取的受众信息包含目标节点时，及时向 Tracker 服务器发送退出传播网络命令。该方法将目标节点从 Tracker 服务器的节点列表缓存中移出，使得后续加入传播网络的节点无法获得目标节点；

（2）占用目标节点空闲资源。向目标节点发起多个连接，模拟多个节点与目标节点之间进行数据传输的现象，直到目标节点的连接数达到最大连接数为止。为了防止被目标节点识别，需要使用虚拟 IP 地址和端口号，为不同实例赋予不同的 IP 地址和端口。同时为了防止 P2P 协议中阻塞算法的干扰，实例与实例之间需要进行一定的数据传输。该方法可以有效阻止新节点与目标节点之间建立连接；

（3）断开与目标节点的现有连接。使用 TCP 连接阻断技术断开可控制网络内节点与目标节点的连接，阻断成功后快速使用仿真客户端向目标节点发起连接，占用目标节点空闲出来的连接资源。此方法可以有效阻断可控制网络内部节点与目标节点之间的 TCP 连接，并且防止新连接的建立。

根据 P2P 节点控制方法设计相应算法来实施控制，控制算法具体步骤如下：

（1）创建待控制目标节点集合 U_{wca} 和已控制目标节点集合 U_{aca}，并对这两个集合进行初始化，根据目标节点选择间隔时间 T_{sit} 设置节点控制间隔时间 T_{cit}；

（2）根据目标节点选择算法，选择需要控制的目标节点集合U_{ytka}，并将其放入集合U_{wca}中，即$U_{\text{wca}}=U_{\text{wca}}\bigcup U_{\text{ytka}}$，并按照节点累积在线时长对节点进行排序；

（3）从集合U_{wca}中取出第一个节点v_1，根据其 IP 地址和端口号，模拟该节点退出 Tracker 服务器的消息，并向 Tracker 服务器发送该消息；

（4）断开网络中与节点v_1相关的现有 P2P 连接；

（5）使用虚拟 IP 地址和端口号生成连接请求消息，并向节点v_1发送该消息，如果节点v_1不可连接，则转到步骤（7）；

（6）如果节点v_1返回 Interested 消息，表示连接建立成功，转到步骤（5）继续建立连接；如果节点v_1返回 Choking 消息，表示节点v_1已阻塞新连接的建立，连接池已满，转到步骤（7）；

（7）从集合U_{wca}中删除节点v_1，即$U_{\text{wca}}=U_{\text{wca}}-\{v_1\}$，并将节点$v_1$加入到已控制节点集合$U_{\text{aca}}$中，即$U_{\text{aca}}=U_{\text{aca}}\bigcup\{v_1\}$；

（8）如果U_{wca}不为空，则转到步骤（3）；否则，获取最新的受众信息，并判断集合U_{aca}中出现在受众信息中的节点列表，并将其存入集合U_{wca}中；

（9）由于 P2P 传播网络的动态性，根据节点控制间隔时间T_{cit}，转到步骤（2），重新选择目标节点，并对目标节点进行控制。

下面对算法所需的系统资源进行分析。

假设N_V为传播网络中的节点数量，p_{ys}为目标节点比率阈值，表示目标节点数最多为$p_{\text{ys}}N_V$。被控特定信息的 Tracker 服务器数量平均为N_{TK}，目标节点的平均连接池数量为N_{LC}，目标节点的平均度值为$<k_{\text{op}}>$。当目标节点数达到$p_{\text{ys}}N_V$时，对目标节点进行控制时需要向 Tracker 服务器发送$p_{\text{ys}}N_V N_{\text{TK}}$条退出消息，与目标节点建立$p_{\text{ys}}N_V N_{\text{LC}}$条连接并进行少量数据传输，断开与目标节点相关的$p_{\text{ys}}N_V<k_{\text{op}}>$条现有 P2P 连接。为了加强目标节点控制效果，最好使用多台计算机组成分布式系统，并在每台计算机上使用多线程技术对目标节点进行控制。

3.5.4　控制策略验证

为了验证不同控制策略对 P2P 特定信息传播网络的控制效果，使用 SEInR 模型对控制效果进行分析。图 3-53 显示了对 SEInR 模型在传播起始阶段使用随机控制和目标控制的效果，图 3-54 显示了对 SEInR 模型在不同传播阶段使用目标控制的效果，图 3-53 与图 3-54 中的控制节点概率都设置为 5%。

从图 3-53 中可以看出，随机控制策略的效果不明显，没有起到抑制特定信息传播的作用；而目标控制策略的效果非常明显，虽然控制节点只占总节点数的 5%，但是对于种

子节点数量来说已经足够。从图 3-54 可以看出，在不同传播阶段使用目标控制时都可以得到良好的控制效果，能够有效地抑制特定信息的传播。

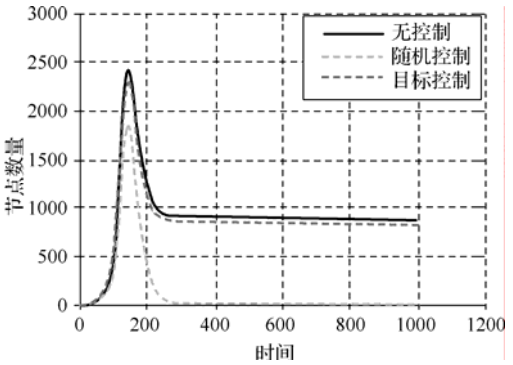


图 3-53　不同控制策略效果比较

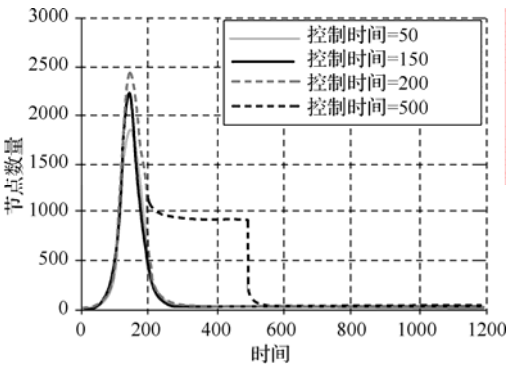


图 3-54　不同传播阶段目标控制效果比较

为了分析控制策略的控制效果，引入控制概率 p，表示从传播网络中移除百分比为 p 的传播节点。同时引入连接率和剩余链路率对传播网络拓扑特性的静态影响进行分析。假设某一时刻，P2P 特定信息传播网络中的节点数量为 N_V，连接数量为 N_E，对于控制概率 p，$0 \leqslant p \leqslant 1$，令 $C(p)$ 为连接率，表示使用控制策略后，仍然相互连接的节点所占百分比。令 $L(p)$ 表示剩余链路率，表示使用控制策略后，仍然存在的边所占比率。因此，$C(p)$ 和 $L(p)$ 的表达式为：

$$\begin{cases} C(p) = \dfrac{c_p}{N_V} \\ L(p) = \dfrac{N_E - e_p}{N_E} \end{cases} \tag{3-57}$$

式中，c_p 为控制后传播网络中相互连接的节点数量，e_p 是被移除边的数量。

为了分析控制策略对实际传播网络拓扑特性的影响，使用 3.4.2 节中的数据进行分析。图 3-55 和图 3-56 分别显示了随机控制策略和目标控制策略在不同控制概率下，连接率 $C(p)$ 和剩余链路率 $L(p)$ 的变化情况。

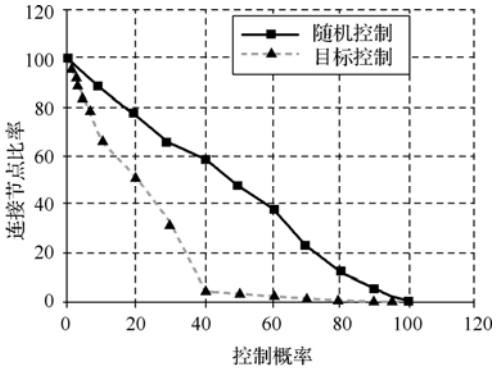


图 3-55　控制概率对连接率的影响

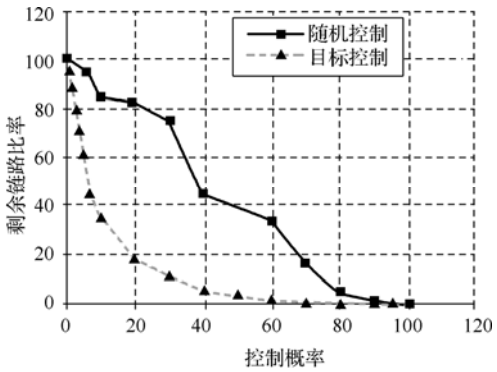


图 3-56　控制概率对剩余链路率的影响

从图 3-55 与图 3-56 可以看出，在目标控制策略下，连接率和剩余链路率都呈现出了快速减小的现象，而在随机控制策略下，连接率是匀速下降的，剩余链路率是随机减少的，一段时间减少的比较慢，另一段时间却减少的很快，主要原因为节点是被随机删除的，度值较大节点被删除的时机不确定，当被删除节点中包含度值较大节点时，剩余链路率减小的速率将会加快。

图 3-57 显示了控制概率设置为 10%，实施两种控制策略后传播网络上的节点数量变化情况。从图 3-57 可以看出，随机控制策略的实际控制效果较差，实时在线节点数量与未控制前相差不大，并且通过对节点的资源拥有情况的分析，发现节点中的资源在不断地

变化，网络中的感染节点也保持在一定水平，说明随机控制策略对 P2P 特定信息传播的抑制效果并不明显。实施目标控制后，传播网络中的在线节点数量变化较大，达到了一定的抑制作用。但是网络中的在线节点数量并没有减少为 0，而是保持在较低的水平。通过对在线节点的状态分析后发现，它们大多数是新增节点，资源拥有率较低、在线时间较短、更新频率较快，说明传播网络已不具备完整的特定信息传播能力，这些节点在等待一段时间后自动放弃特定信息下载，转而寻求其他获取方式。通过上述分析说明，目标控制策略能够明显地抑制 P2P 特定信息的传播。

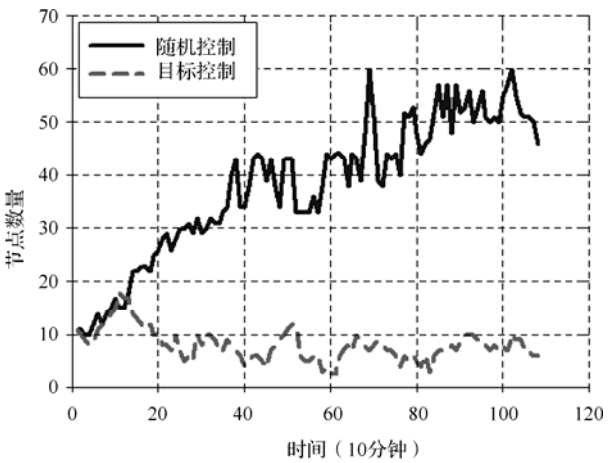


图 3-57　不同控制策略对传播网络的动态控制效果比较

在目标控制策略实施一段时间后，种子节点很少与其他节点建立连接并传输数据，在获取的受众信息中也没有发现种子节点的 IP 地址。说明设立的种子节点已从传播网络中被有效移出，使得 P2P 特定信息传播网络得到有效的控制。

综上所述，P2P 特定信息传播网络控制策略采用免疫策略思想，以 SEInR 模型为基础，通过分析对比随机控制策略与目标控制策略的控制效果，选择了目标控制策略作为主要控制手段。根据 P2P 特定信息传播规律和特点，采用 P2P 节点控制方法，实现对 P2P 特定信息传播网络的有效控制，仿真和实验结果表明，达到了较好的控制效果。

参考文献

[1] J. Liang, R. Kumar, Et Al. Pollution in P2P File Sharing Systems[C]. INFOCOM 2005. Miami, Florida, USA, 2005.

[2] 俞嘉地. BitTorrent 对等网文件共享系统关键技术研究[D]. 上海交通大学博士学位论文, 2007.

[3] Deke Guo, Jie Wu, Honghui Chen, et al. Theory and network applications of dynamic bloom filters[C]. Proceedings of the 25th Annual Joint Conference of the IEEE Computer and Communications Societies, 2006: 1-10.

[4] Andrei Broder, Michael Mitzenmacher. Network applications of bloom filters: A survey[J]. Internet Mathematics, 2004, 1(4): 485-509.

[5] 刘浩. P2P 网络的若干关键问题研究[D]. 华南理工大学博士学位论文, 2010.

[6] M. E. J. Newman. Power law, Pareto distributions and ZIPf's law[J]. Contemporary Physics, 2005, 46(5): 323-351.

[7] Ramon Ferrer I. Cancho, Ricard V. Solé. Optimization in complex networks[J]. Lecture Notes in Physics, 2003, 625: 114-126.

[8] 王林. 复杂网络的 SCALE-FREE 性、SCALE-FREE 现象及其控制[D]. 西北工业大学博士学位论文, 2006.

[9] 陈希孺. 基尼系数及其估计[J]. 统计研究, 2004, 21(8): 58-60.

[10] M. E. J. Newman. Assortative mixing in networks[J]. Physical Review Letters, 2002, 89(20): 208701.

[11] Paolo Crucitti, Vito Latora, Massimo Marchiori, et al. Error and attack tolerance of complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2004, 340(1): 388-394.

[12] Reuven Cohen, Keren Erez, Daniel B. Avraham, et al. Resilience of the internet to random breakdowns[J]. Physical Review Letters, 2000, 85(21): 4626-4629.

第4章

社交网络用户关系模型

4.1 引言

在 1.3 节中，对在线社交网络（简称社交网络）及其信息传播模式进行了介绍。从中可以看出，互联网中的社交网络极大地方便和丰富了人们的社会交往和信息交流，受到广大网民的欢迎和热捧，使社交网络成为一种非常流行的新型互联网应用。同时，社交网络也带来了信息安全方面的挑战，例如利用社交网络平台进行非法联络、传播谣言、煽动闹事等，容易引起社会群体事件。

这种新型的互联网应用同时也引起国内外研究人员的极大关注，从不同的角度对社交网络进行了研究，包括社交网络的基本特性、网络结构、信息传播、用户关系、连接强度等，通过建立相应的数学模型，对社交网络及其信息传播特性进行分析，找出其中的内在规律，为优化社交网络结构，改善社交网站服务，正确引导网络舆论等提供科学依据。

本章主要对社交网络的用户关系、连接强度以及对信息传播和用户行为的影响等问题进行分析和研究，有助于认识社交网络信息交流和互动的内在动力和规律。

4.2 社交网络信息传播模型

针对各种社交网络的信息传播问题，人们提出了一些信息传播模型，主要有经典信息传播模型、巴斯扩散模型及谣言传播模型等。

4.2.1 经典信息传播模型

文献[1]在研究社交网络影响力最大化问题时提出了三个经典的信息传播模型：独立级联模型、带权级联模型和线性阈值模型。

1．独立级联模型

在独立级联模型中，社交网络被抽象成为一个无向图 $G = (V, E)$，其中 V 代表网络中的用户，E 代表用户间的关系。当信息在网络中传播时，节点有两种状态：活跃和非活跃，其中活跃表示节点接受信息，而非活跃则表示节点还未接受信息，并且节点只能由非活跃转为活跃状态，而不能由活跃转为非活跃状态。信息只能通过边 E 传播，没有边相连的节点则不能直接相互影响。每个活跃节点对非活跃节点的影响是相互独立的，活跃节点以固定的概率 p 激活非活跃节点，其中 $0<p<1$。

2．带权级联模型

带权级联模型与独立级联模型非常相似，不同之处在于带权级联模型是一个有向图，邻居节点间的影响因子不对称，即 $P_{uv} \neq P_{vu}$，且节点的影响因子为 $1/d$，其中 d 为节点的度大小。

3．线性阈值模型

线性阈值模型也是把社交网络抽象成一个图 $G = (V, E)$，其中 V 代表网络中的用户，E 代表用户间的关系。在线性阈值模型中，节点 v 与其相邻节点 w 的边都有一个权值 b_{vw}，且 $b_{vw} \neq b_{wv}$，节点 v 所有相邻节点权值之和小于或等于 1，每个节点 v 都随机给出一个阈值 $\theta_v(0 \leqslant \theta_v \leqslant 1)$。而非活跃节点 v 被激活条件是其相邻所有活跃节点的权值之和大于或等于阈值 θ_v。

4.2.2　巴斯扩散模型

巴斯扩散模型（Bass Diffusion Model）及其扩展理论是由 Frank M. Bass 等[2]提出的，主要针对创新性产品信息通过社交网络进行传播及预测问题。

巴斯扩散模型将一项创新性产品信息在市场上的扩散归结为两大因素的影响，一是创新的或外部影响，这种影响主要依靠大众媒介（如广告）来扩散；二是模仿的或内部影响，它是指人与人之间（即已采用者对未采用者）的口头交流影响，也就是口碑相传方式。巴斯扩散模型如下：

$$N_t = N_{t-1} + p(m - N_{t-1}) + q\frac{N_{t-1}}{m}(m - N_{t-1}) \qquad (4\text{-}1)$$

式中，m 为潜力系数，即潜在使用者总数，p 为创新系数，即外部影响，尚未使用该产品的人，受到大众传媒或其他外部因素的影响，开始使用该产品的可能性；q 为模仿系数，即内部影响，尚未使用该产品的人，受到使用者的口碑影响，开始使用该产品的可能性。

4.2.3 谣言传播模型

文献[3，4]采用 SIR 传播动力模型来描述谣言在社交网络中的传播过程，在 SIR 模型中，感染者以概率 v 把传染病传给易感者，易感者被感染后成为新的传染源；感染者以概率 δ 被治愈，治愈者对疾病具有免疫能力。谣言传播过程与疾病传播过程相类似，随着谣言在人群中的扩散，人群最终分化为两类人群，一类是听过谣言而产生免疫的治愈人群，另一类是未听过谣言而被感染的易感人群。当治愈人群处于比较小的数值区域时，治愈人群与易感人群、感染人群和谣言消亡时间服从幂律分布。

关于 SIR 传播动力模型介绍参见 2.5 节。

4.3 社交网络关系强度模型

社交网络是通过用户之间的信息交流和互动行为而形成的一种网络结构，用户关系是社交网络的基本特性之一，也是人们重点研究的对象。

4.3.1 用户关系特性

1．强连接与弱连接

人们是如何找到他们所需要的工作呢?是靠亲朋好友的帮忙还是通过各种招聘广告或招聘会？20 世纪 60 年代末，哈佛大学的 Mark Granovetter 对这些问题进行了研究，他在波士顿地区走访了近 100 个人，并问卷调查了近 200 个人。

Granovetter 发现，在寻找工作过程中，那些关系紧密的朋友反倒没有那些关系一般的甚至只是偶尔见面的朋友更能发挥作用。事实上，关系紧密的朋友也许根本帮不上忙。在 Granovetter 的论文“弱连接的强度”（Strength of Weak Ties）[5]中给出了如下的例子：

Edward 在高中时，他认识的一个女孩邀请他参加了一个聚会。在聚会上，Edward 遇到了比他大十岁的那个女孩姐姐的男朋友。三年之后，当 Edward 辞去了工作之后，在当地的住所遇到了这位一面之缘的朋友。在交谈中，这个朋友说起他所在公司现在需要一个制图员，于是 Edward 申请了这个工作，并顺利地被雇佣。

Granovetter 这篇论文最初被《美国社会科学评论》杂志拒之门外，直到 4 年之后，才被《美国社会学》杂志接受。现在它已被公认为是现代社会学最有影响的论文之一。

Granovetter 指出，在传统社会中每个人接触最频繁的是自己的亲人、同学、朋友、以及同事等。这是一种十分稳定但是传播范围有限的社会认知，是一种强连接（Strong Ties）关系。同时，还存在另外一类相对于强连接关系更为广泛的，却是肤浅的社会认

知，例如一个被无意间提到或者从收音机偶然听到的一个人，这种连接关系称为弱连接（Weak Ties）关系。关于强弱连接的界定，Granovetter 设计了四个指标，它们分别是互动时间、情感强度、亲密程度以及互惠行动的内涵，但没有给出用来测量和判别强弱连接的标准，Granovetter 采用互动次数来测量和判别连接的强度。

很自然地，人们都会想，强连接可能比弱连接更有用处。然而，Granovetter 调查发现，就寻找工作而言，弱连接的机会要比强连接高得多。因此，Granovetter 认为在探究一些网络现象时，使用弱连接的概念比使用强连接的概念要重要得多。

强连接关系通常表示行动者彼此之间具有高度的互动，在某些存在的互动关系形态上较为亲密。因此，通过强连接传播的信息通常是重复的，容易自成一个封闭的系统。网络内的成员由于具有相似的态度、高度的互动频率通常会强化原本认知的观点，而降低了与其他观点的融合，强连接网络并不是一个可以提供创新机会的渠道。

相对于强连接关系，弱连接则能够在不同的团体间传递非重复性的信息，使得网络中的成员能够增加修正原先观点的机会。事实上，在信息扩散和传播方面，弱连接起着同样的作用。一个人的亲朋好友圈子里的人可能相互认识，在这样的圈子中，他人提供的交流信息总是冗余的。例如，从这个朋友或亲戚听到的信息，可能早已经从另一个朋友那里听说了，而他们之间也都相互交谈过此话题。日常生活中不乏这样的事例。

2．服从权威现象

对于人类社会是否存在服从权威现象，耶鲁大学 Stanley Milgram 设计了著名的权威服从实验[6]。Milgram 想知道，当权威人物命令一个人去伤害他人的时候，这个人究竟会残酷到何种地步。在第二次世界大战之后，人们都想知道，人类如何被激发对同类犯下如此残酷的罪行，不仅是那些武装部队，就连普通人都被强迫去实施最为残酷可怕的暴行。然而，Milgram 没有去调查战争中的极端情况，而是在实验室相对正常的环境下观察人们的反应，即当一个人被要求向另一个人实施电击的时候会有什么表现？人们会无视自己的忧虑而服从命令到什么样的程度？

实验结果表明，在参与试验的人群中没有一个人是虐待狂，甚至没有任何人格上的缺陷，但是他们会服从权威的命令做出一些违背道德伦理的事情，不是因为他们具有服从性的人格，而是当时的权威服从暗示的情景所致。因此，Milgram 认为人类比想象中更加愿意服从权威（Obedience to Authority），服从权威是“一个根深蒂固的倾向，能够压倒道德、同情以及社会秩序的冲动”。

在社交网络中，信息传播也存在服从权威现象，例如在一条普通用户连接超级明星的关注边中，一般用户很容易受到超级明星的影响而转发他们的信息。

4.3.2 关系强度估计

在社交网络中，用户之间存在着强连接或弱连接关系。对于不同类型的社交网络，采用不同的测量和判别方法来评估用户关系的强弱。对于互联网中的社交网络，通常采用关系强度来估计用户关系的强弱。

社交网络的用户关系强度估计具有现实意义，可以用来改善和优化社交网络的多个方面性能，如推荐、搜索、定位等。

（1）连接推荐。在新浪微博、人人网等社交网络中，系统自动为用户推荐新的连接，通过用户对之间的关系强度估计，如地理位置接近，兴趣相同等，能够很容易地把关系强度较高的人推荐给用户。

（2）项目推荐。关系强度估计可以用于改善社交网络为用户提供的推荐服务，如工作推荐、购物推荐等，因为一个人的偏好和选择更有可能接近于和他密切相关的人。例如，在工作推荐网站上，一个用户想加入一个工作推荐组或者想阅读一些新闻文章，向该用户推荐与他关系强度高的人群的相关活动，用户采用的成功率会更高。

（3）新闻推送。新闻推送是社交网络的一个重要功能，能够把实时更新、状态变化、新活动、新职位或联系人变化推送给用户。通过关系强度估计，可以为一个在线成员构建个性化新闻，优先更新与用户之间具有强连接关系的信息内容，甚至可以去除或减弱伪造联系人更新的困扰。

（4）人员搜索。根据查询者和查询结果的关系强度进行搜索结果排名，用户可以更迅速地找到他想要找的人。

（5）内容定位。通过关系强度估计，突出显示社交网络中与用户关系强度高的人员或内容，方便用户迅速找到感兴趣的内容。

在社交网络中，不仅存在着强连接和弱连接这两种极端关系，还存在着处于两者之间的用户关系，因此通过关系强度能够准确地刻画社交网络中的各种用户关系。然而，在实际的社交网络中，由于建立朋友关系成本很低，结果造成社交网络中只包含强连接和弱连接这两种极端关系，而很少或没有提供用于区分其他用户关系的信息，也就难以对关系强度进行准确的估计。

通常，在社交网络中不仅有用户关系信息，还包含有用户之间的辅助交互信息，利用这些信息可以建立关系强度模型。事实上，一般的社交网络都提供用于加强信息传输和促进社区维护的基础信息，在社交网络系统中保存有用户的底层交互记录，可以用来识别成员之间联系的紧密程度。例如，在新浪微博用户中，每个用户有一个页面用来展示他们的信息，朋友们可以在上面留言。特殊的用户可能有成千上万的朋友，而由于精力有限或者资源有限，他们很可能只与亲密的朋友进行交流。

下面给出一个潜变量模型，利用用户配置文件相似度和用户交互活动特征来估计关系强度，自动识别出社交网络用户关系的强弱程度。

1．模型定义

潜变量模型是利用用户交互活动数据和用户配置文件属性数据来估计关系强度的。由于用户关系强度隐含在用户配置文件属性相似度之内，而相似度又引起用户之间的交互活动，因此该模型充分利用了这两种不同类型信息的特点，能够捕捉到隐含在社交活动中的因果关系。

另外，由于用户配置文件数据具有相对全面、稳定和容易获取等特点，而用户交互数据通常是稀疏的，并具有时间性，因此需要区别对待这两种数据，在判别模型中以配置文件属性相似度为特征，而在生成模型中以用户交互活动为特征。这样，潜变量模型具有两个优点：基于用户配置文件数据的判别模型能够准确地估计用户关系强度，基于用户交互数据的生成模型能够灵活地处理交互数据易丢失的问题。

潜变量模型是建立在社会学同质性理论[7]的基础上。所谓同质性是指人们更倾向于与其具有类似特征的人建立关系，即具有与同类人来往的倾向性，关系越强相似度越高。因此，以社交网络节点配置文件相似度为特征来表征关系强度是合理的。节点配置文件中的学校、公司、地理位置、用户之间关注状态等属性可以用来推导出节点相似性。

由于每个用户只拥有有限数量的资源（例如时间）来建立和维护关系，所以他们很可能把这些资源投向他们认为更重要的关系上。这些交互活动可能包括查看对方用户信息、连接关注、转发、评论等。用户对之间的关系强度越强，某种类型的交互活动也就越有可能发生。可以合理地假设关系强度直接影响用户对之间交互活动的性质和频率，将关系强度作为用户交互活动的隐含原因，以关系强度为条件，交互活动变量之间彼此相互独立。

潜变量模型定义如下：

$$P(f^{(ab)}, m^{(ab)} \mid v^{(a)}, v^{(b)}) = P(f^{(ab)} \mid v^{(a)}, v^{(b)}) \prod_{i=1}^{n} P(m_i^{(ab)}) f^{(ab)} \tag{4-2}$$

式中，$v^{(a)}$为用户 a 的属性向量，$v^{(b)}$为用户 b 的属性向量，$m_i^{(ab)}$ 为用户 a 和 b 之间所发生的 n 种交互活动中的一个，$i=1,2,\cdots,n$，$f^{(ab)}$ 为用户 a 和 b 之间的关系强度。在模型中，$f^{(ab)}$ 受 $v^{(a)}$和 $v^{(b)}$影响，而 $m_i^{(ab)}$ 受 $f^{(ab)}$ 影响。

虽然关系强度变量 f 代表了用户对之间的相似性和交互性，但是变量值难以从数据中直接观察得到，所以将模型中的变量 f 作为潜变量来对待。模型既可用于估计有向关系强度又可用于估计无向关系强度，取决于如何指定用户对之间的属性相似性和交互性。这里主要研究有向关系强度估计问题，估计值 $f^{(ab)}$ 有可能不等于 $f^{(ba)}$，因为交互活动是有方向性的，用户 a 转发用户 b 的信息，并不代表用户 b 也转发了用户 a 的信息。

2．模型实例化

潜变量模型可以采用合适的方法进行实例化，这主要取决于属性和交互性的可解释性。这里采用广泛使用的高斯分布来模型化关系强度的条件概率，关系强度由属性相似度决定。

定义 $s_i(v^{(a)},v^{(b)})$，$i=1,2,\cdots,n$ 为节点对 a、b 之间相似度测量的数据集，则 $f^{(ab)}$ 和 $v^{(a)}$、$v^{(b)}$之间的依赖可以表示如下：

$$P(f^{(ab)}\mid v^{(a)},v^{(b)})=N(w^{\mathrm{T}}s(v^{(a)},v^{(b)}),u) \tag{4-3}$$

式中，s 为 $v^{(a)}$和 $v^{(b)}$计算出的相似度向量，w 为估计出的 n 维权重向量，u 为高斯分布模型的方差，实验中 u 取值 0.5。

图 4-1 为模型具体实例描述的图形表示，其中用 $s^{(ab)}$代替了 $v^{(a)}$和 $v^{(b)}$。在模型中，由 $f^{(ab)}$ 条件独立地给出每个 $m_i^{(ab)}$ 的概率分布。这里定义的所有交互活动为二值变量，例如用户 a 转发了用户 b 的信息则此变量为 1，否则为 0。

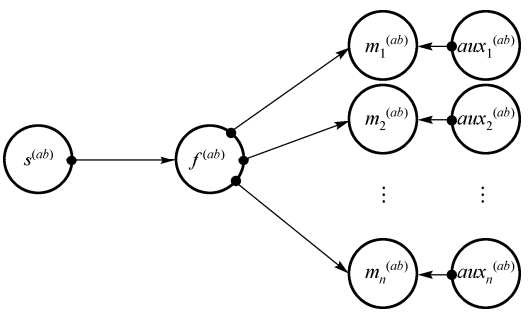


图 4-1　模型具体实例描述的图形表示

另外，为每个交互活动 m 引入辅助变量 $aux_{i1}^{(ab)},aux_{i2}^{(ab)},\cdots,aux_{il_i}^{(ab)}$ 来增加模型的精确性。这些变量用来表示引起交互活动的辅助原因，它们独立于关系强度。例如，一个用户的关注总人数可以代表该用户参与社交网络的活跃程度，可以根据关注总人数的多少来调整被该用户关注而带来的对关系强度的影响。

用一个逻辑函数来形式化由 $f^{(ab)}$ 和 $aux_i^{(ab)}$ 给出的 $m_i^{(ab)}$ 的条件概率，即：

$$P(m_i^{(ab)}=1\mid f^{(ab)},aux_i^{(ab)})=\frac{1}{1+\mathrm{e}^{-(\theta_{i1}aux_{i1}^{(ab)}+\theta_{i2}aux_{i2}^{(ab)}+\cdots+\theta_{il}aux_{il}^{(ab)}+\theta_{il+1}f^{(ab)}+\beta)}} \tag{4-4}$$

式中，θ_i 为估计的参数集合，$\theta_i=[\theta_{i1},\theta_{i2},\cdots,\theta_{il},\theta_{il+1}]^{\mathrm{T}}$，简化表示为：

$$P(m_i^{(ab)}=1\mid d_i^{(ab)})=\frac{1}{1+\mathrm{e}^{-(\theta_i^{\mathrm{T}}d_i^{(ab)}+\beta)}} \tag{4-5}$$

式中，$d_i^{(ab)}$ 为辅助变量与关系强度的向量，$d_i^{(ab)}=[aux_i^{(ab)},f^{(ab)}]^{\mathrm{T}}$。

通常在不增加推导难度的情况下，可以采用合适的广义线性模型来定义交互活动变量。可以使用泊松回归将交互活动表示为计数数据，计数数据是将观察单位按某种类别或属性进行分组，清点各组观察单位所得的数据量。为了避免过度拟合，对参数 w 和 θ 进行多元常态调整，得到高斯先验函数，即：

$$P(w) \propto \mathrm{e}^{-\frac{\lambda_\omega}{2} w^{\mathrm{T}} w} \tag{4-6}$$

$$P(\theta_i) \propto \mathrm{e}^{-\frac{\lambda_\omega}{2} \theta_i^{\mathrm{T}} \theta_i}, \quad i = 1,2,\cdots,n \tag{4-7}$$

数据为 N 个用户对样本，表示为 $D = \{(a_1,b),(a_2,b_2),\cdots,(a_N,b_N)\}$。

在训练过程中，变量 $v^{(ab)}$，$m^{(ab)}$ 和 $aux_i^{(ab)}$，$((ab) \in D, i = 1,2,\cdots,n)$ 都是可见的。因为属性相似度是在用户 v 的基础上预先计算的。为简化公式，根据公式（4-2）得到如下的联合概率：

$$\begin{aligned} & P(D \mid w,\theta) P(w,\theta) \\ & \left(\prod_{(a,b) \in D} P(f^{(ab)}, m^{(ab)} \mid v^{(a)}, v^{(b)}, w, \theta) \right) P(w) P(\theta) \\ & = \prod_{(a,b) \in D} P(f^{(ab)} \mid v^{(a)}, v^{(b)}, w) \prod_{c=1}^{n} \prod P(m_i^{(ab)} \mid f^{(ab)}, \theta_i) P(w) P(\theta_i) \\ & \propto \prod_{(a,b) \in D} \left(\mathrm{e}^{-\frac{1}{2v}(w^{\mathrm{T}} s^{(ab)} - f^{(ab)})^2} \prod_{i=1}^{n} \frac{\mathrm{e}^{-(\theta_i^{\mathrm{T}} d^{(ab)} + \beta)(1 - m_i^{(ab)})}}{1 + \mathrm{e}^{-(\theta_i^{\mathrm{T}} d^{(ab)} + \beta)}} \right) \cdot \mathrm{e}^{-\frac{\lambda_\omega}{2} w^{\mathrm{T}} w} \prod_{i=1}^{n} \mathrm{e}^{-\frac{\lambda_\theta}{2} \theta_i^{\mathrm{T}} \theta_i} \end{aligned} \tag{4-8}$$

式中，$s^{(ab)}$ 为所有观测变量，$s^{(ab)} = s(v^{(a)}, v^{(b)})$。

潜变量模型可以采用两种不同的估计方法，第一种方法是先推导潜变量 f 的分布，找到参数 $\hat{w}$ 和 $\hat{\theta}$ 的点估计，求极大联合似然估计 $P(m, \hat{w}, \hat{\theta} \mid v)$，这种方法一般采用期望最大化（EM）算法。第二种方法是将潜变量 f 看作一个名义参数，找到点估计 $\hat{w}$ 和 $\hat{\theta}$，$\hat{f}$，求极大似然估计 $P(m, \hat{f}, \hat{w}, \theta \mid v)$。由于潜变量 f 在 EM 的迭代过程中难以管理，因此这里采用第二种估计方法。

对公式（4-8）取对数，得到潜变量的对数似然函数，即：

$$\begin{aligned} L(f^{(\{(a,b) \in D\})}, w, \theta_t) = & \sum_{(a,b) \in D} -\frac{1}{2u} (w^{\mathrm{T}} s^{(ab)} - f^{(ab)})^2 \\ & + \sum_{(a,b) \in D} \sum_{i=1}^{n} -(1 - m_i^{(ab)})(\theta_i^{\mathrm{T}} d_i^{(ab)} + \beta) - \log(1 + \mathrm{e}^{-(\theta_i^{\mathrm{T}} d_i^{(ab)} + \beta)}) \\ & - \frac{\lambda_w}{2} w^{\mathrm{T}} w - \sum_{i=1}^{n} \frac{\lambda_\theta}{2} \theta_i^{\mathrm{T}} \theta_i + C \end{aligned} \tag{4-9}$$

在公式（4-9）中包括的二次项和对数逻辑函数都是凹函数，因此函数 L 是凹函数。下面采用梯度法对参数 w 和 θ_i $(i=1,2,\cdots,n)$ 以及潜变量 $f^{(ab)}$ $(a,b)\in D$ 进行优化，以便找到 L 的最大值，通过优化得到如下结果：

$$\frac{\partial L}{\partial f^{(ab)}}=\frac{1}{u}(w^{\mathrm{T}}v^{(ab)}-f^{(ab)})+\sum_{i=1}^{n}\left(m_i^{(ab)}-\frac{1}{1+\mathrm{e}^{-(\theta_i^{\mathrm{T}}d_i^{(ab)}+\beta)}}\right)\theta_{i,l_i+1} \tag{4-10}$$

$$\frac{\partial L}{\partial \theta_i}=\sum_{(ab)\in D}\left(m_i^{(ab)}-\frac{1}{1+\mathrm{e}^{-(\theta_i^{\mathrm{T}}d_i^{(ab)}+\beta)}}\right)d_i^{(ab)}-\lambda_\theta\theta_i \tag{4-11}$$

$$\frac{\partial L}{\partial w}=\frac{1}{u}\sum_{(ab)\in D}(f^{(ab)}-w^{\mathrm{T}}s^{(ab)})s^{(ab)}-\lambda_w w \tag{4-12}$$

采用牛顿-拉普森优化方案迭代更新 $f^{(ab)}$ 和 θ_i，直至收敛。对于 w，采用岭回归解析公式（4-12）的根。

$$f^{(ab)new}=f^{(ab)old}-\frac{\partial L}{\partial f^{(ab)}}\Big/\frac{\partial^2 L}{\partial (f^{(ab)})^2} \tag{4-13}$$

$$\theta_i^{new}=\theta_i^{old}-\frac{\partial L}{\partial \theta_i}\Big/\frac{\partial^2 L}{\partial \theta_i \partial \theta_i^{\mathrm{T}}} \tag{4-14}$$

$$w^{new}=(\lambda_w I+s^{\mathrm{T}}s)^{-1}s^{\mathrm{T}}f \tag{4-15}$$

式中，s 为观测变量向量，$s=[s^{(a_1b_1)}s^{(a_2b_2)}\cdots s^{(a_Nb_N)}]^{\mathrm{T}}$；$f$ 为关系强度向量，$f=[f^{(a_1b_1)}f^{(a_2b_2)}\cdots f^{(a_Nb_N)}]^{\mathrm{T}}$。

整个优化过程如算法 4-1。

算法 4-1　学习算法

重复步骤（1）~（3）直至收敛：

（1）牛顿-拉普森方案迭代每一步：按照式（4-13）更新 $f^{(ab)}$;

（2）牛顿-拉普森方案迭代每一步：按照式（4-14）更新 θ_i;

（3）按照式（4-15）更新 w。

对于一个新用户对 (a,b)，学习模型有如下两种使用方法。

（1）如果两个用户的属性 $v^{(a)}$、$v^{(b)}$ 和用户对的交互活动 $m_1^{(ab)},\cdots,m_i^{(ab)}$ 已知，可以按照学习算法步骤（1）来估计关系强度 $f^{(ab)}$；

（2）当无法得到交互数据时，只需要通过公式（4-13）来推断关系强度 $f^{(ab)}$。

由于交互数据经常是稀疏的、有时间性的，并且难以获得，所以对于社交网络来说第

二种方法更常用一些。这个混合模型的优势就在于模型的下半部分是生成模型，所以整个模型在训练过程中不会遇到太多的交互数据丢失问题。一旦模型完成学习，对于新数据只需要使用混合模型的上半部分就可以推断出潜变量。另外，模型的生成模型部分也可以用于预测未来的交互活动。

4.3.3　模型验证

下面通过实验数据对社交网络用户关系强度估计模型进行测试和验证。

1．实验数据集

实验数据来源于新浪微博。在新浪微博中，用户可以快速地发布 140 字以内的信息，并被用户关注者所转发。用户可以关注任何感兴趣的用户，整个用户的关注或被关注关系构建一个有向网络图。在该网络图中，每一条边都是信息传播的桥梁。微博中的用户不仅可以自由选择感兴趣的用户，而且可以发布微博，与其他用户互动。用户的互动可以通过邮件、评论、转发等方式。互动性是微博与传统媒体的最大区别。

在用户关系强度估计中，需要综合考虑兴趣、共同关注、距离以及性别等特征。实验中随机选择 100 个微博用户作为种子节点，从这些节点开始采样。通过选择连接的或非连接的节点对，将每个种子节点和它所有的邻居添加到连接图。从这些直接和间接的邻居中采样大约 100000 对数据。每个样品对包括一个种子节点和一个用户两跳之内的邻居节点。对于每一对用户 (a,b)，通过计算若干特征来推导用户配置文件之间的相似度。定义总体相似度为：$s=[s^{(a_1b_1)}s^{(a_2b_2)}\cdots s_7^{(a_Nb_N)}]^{\mathrm{T}}$。7 个特征描述如表 4-1 所示。

表 4-1　新浪微博用户配置文件特征

特征	描述
s_1	如果 a 和 b 曾在同一所学校则为 1，否则为 0
s_2	如果 a 和 b 在同一座城市则为 1，否则为 0
s_3	如果 a 和 b 在同一个省份则为 1，否则为 0
s_4	如果 a 和 b 性别相同则为 1，否则为 0
s_5	如果 a 和 b 之间有兴趣范围相同则为 1，否则为 0
s_6	如果 a 和 b 之间相互关注则为 1，否则为 0
s_7	a 和 b 之间共同关注数的对数

除了相似度特征，还需要在模型中考虑各种类型的用户交互活动特征。4 种交互活动是：连接、转发、评论和邮件，其特征分别用 $m_1^{(ab)}$、$m_2^{(ab)}$、$m_3^{(ab)}$ 和 $m_4^{(ab)}$ 来表示，如表 4-2 所示。

表 4-2　新浪微博用户交互特征

特征	描述
m_1	如果 a 和 b 之间有连接则为 1，否则为 0
m_2	如果 a 和 b 之间有转发行为则为 1，否则为 0
m_3	如果 a 和 b 之间有评论行为则为 1，否则为 0
m_4	如果 a 和 b 之间有邮件行为则为 1，否则为 0

模型中每种类型的交互活动包含一个辅助变量，用来表示与用户 a 存在指定方式交互活动的总人数。例如 m_1 的辅助变量就表示与节点 a 建立连接的总节点数。

2．模型有效性验证

现在使用潜变量模型来估计实验数据中的用户对之间的关系强度，其估计结果可应用于不同的推荐任务中。微博的一个重要任务是将特定的人推荐给用户。例如，当一个用户寻找自己感兴趣的信息时，系统可以通过估计关系强度将与该用户关系强度较高的用户和信息推荐给他。

在实验中，首先将地理位置、兴趣范围、性别等属性特征用于模型学习和关系强度估计，然后对估计出来的用户对之间的地理位置、兴趣范围、性别等属性进行测量。在测量时，首先按照不同方式对用户对进行排名，然后测量不同排名方式用户对的 AUC （受试者工作特征曲线下面积）。

用户对的排名方式有如下几种。

（1）评论连接：按照一方评论另一方信息的数量对用户对进行排名。

（2）转发连接：按照一方转发另一方信息的数量对用户对进行排名。

（3）双向关注连接：按照双方是否互相关注对用户对进行排名。

（4）单向关注连接：按照一方转发另一方信息的数量对用户对进行排名。

（5）交互活动数量：按照所有类型交互活动的数量对用户对进行排名。

（6）配置文件相似性：根据用户对之间的相似度 $\sum_k s_k^{(ab)}$ 进行排名。

（7）关系强度：根据关系强度计算结果进行排名。

方式（1）~（4）对应着数据中不同类型的连接，方式（5）和（6）分别代表交互活动和相似性的应用，方式（7）对应关系强度计算结果。这些结果分别使用了模型上半部分（配置文件相似度）和模型下半部分（交互活动），排名结果如图 4-2 所示，从图 4-2 可以看出，基于关系强度的排名方式比其他排名方式要高出很多，这说明模型能够很好地将配置文件相似性和交互活动结合起来自动识别出相似用户对。

下面使用历史数据对关系强度估计性能进行评估。首先在模型中每次使用一个特征对关系强度进行估计，然后测量估计出的关系强度与历史数据的对应程度。按照关系强度估

计值对用户对进行排名，同时使用一个双向关注变量计算曲线下面积 AUC。双向关注变量设定为：如果用户对之间存在双向关注，则变量值为 1，否则为 0。

评估结果如图 4-3 所示，从图 4-3 可以看出，基于关系强度的排名方式明显优于其他方式，这表明了采用潜变量模型进行关系强度估计能够更容易地识别出与用户相关的人，并且可以为用户提供更具相关性的推荐。同时也证明了这种关系强度建模方法与人们之间关系的感知方式是比较接近的。

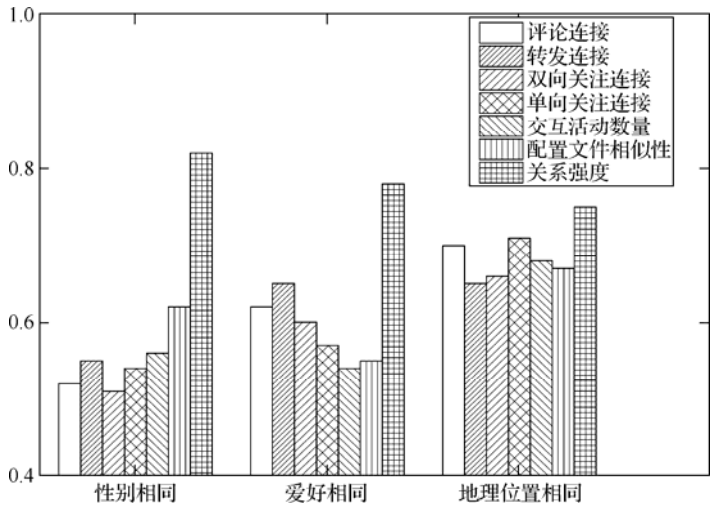


图 4-2　用户对排名方式的 AUC 结果对比

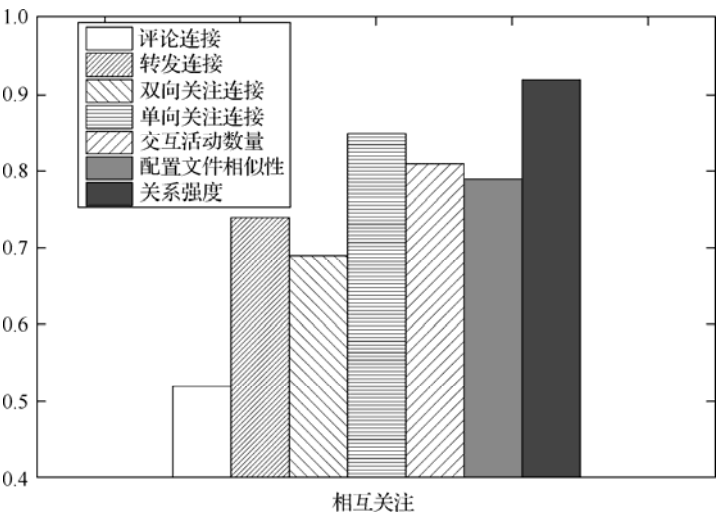


图 4-3　关系强度估计性能评价结果

4.4 社交网络弱连接分析

社交网络作为一种新兴的社会媒介，在社会信息传播过程中扮演着重要的角色。然而由于社交网络结构的复杂性，社交网络的信息传播规律和传播机制与其他类型的网络完全不同，人们对社交网络的用户关系、网络结构以及信息传播之间的内在联系缺乏了解和认知。

在社交网络的用户关系中明显存在着强连接和弱连接关系，通常采用关系强度来测量和估计用户连接的强弱。从信息传播的角度，弱连接虽然不如强连接那样坚固，却有着极快的、低成本和高效能的传播效率，因此对弱连接的分析和研究是非常重要的。

下面主要研究弱连接对社交网络结构和信息传播的影响。

4.4.1 连接强度模型

Onnela 等[8]给出了一种通过计算两个节点之间的邻居节点重叠率来估计连接强度的方法，节点 i 和节点 j 之间的邻居节点重叠率定义如下：

$$w_{ij}=\frac{c_{ij}}{k_i-1+k_j-1-c_{ij}} \tag{4-16}$$

式中，c_{ij} 为节点 i 和节点 j 共同的邻居数，k_i 为节点 i 的邻居数，k_j 为节点 j 的邻居数，w_{ij} 为两端分别为节点 i 和节点 j 的边的权值。

依据该方法得出的权值越低，其相应的连接强度也就越低。根据弱连接理论，如果节点 i 和节点 j 之间的连接强度高，那么由于两个节点经常联系并且具有许多相同的属性，两个节点的共同朋友也会多，相应边的权值就会大。因此，该权值能够在一定程度上反映了弱连接的强度，称为朋友重叠率指标。

然而，该指标也存在一定的问题，例如，在图 4-4 和图 4-5 中，如果利用朋友重叠率指标来计算，其连接强度均为 0。然而，通过人工分析其结果未必是正确的。

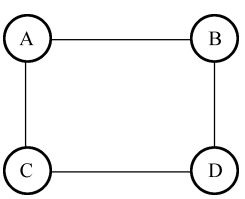


图 4-4 朋友重叠率指标问题示例 1

在图 4-4 中，边 AB 的强度为 0，因为 A 与 B 没有共同的朋友。虽然 A 与 B 没有共同的朋友，但是 A 的朋友 C 与 B 的朋友 D 也相互为朋友，显然 A 与 B 的关系强度并非

为 0，A、B、C、D 四个节点之间的关系联系紧密。因此，在这种情况下，朋友重叠率指标存在片面强调朋友重叠率而忽略了两个朋友圈间关系的问题。

在图 4-5 中，节点 A 有 B、C、D、E、F、G 等众多的朋友，然而 B 节点只有两个朋友 A 和 G，通过朋友重叠率指标计算其连接强度为 0.25。然而，人工分析可知，作为 B 节点的两个朋友之一 A，其连接强度是很高的。从信息阻断的角度，当信息在 A 节点传播时，移去边 AB 最好情况可以使传播范围减 1，这样的边并不足以当作弱连接。以此类推，当 k_i 与 k_j 的值相差悬殊时，必然导致 c_{ij} 与 $k_i-1+k_j-1-c_{ij}$ 的值相差悬殊，从而使结果偏向弱连接。因此，在这种情况下，朋友覆盖率指标存在因节点本身朋友数量相差悬殊而产生的较大误差。

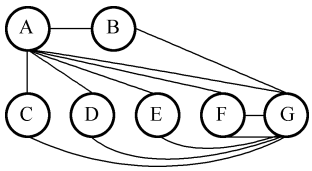


图 4-5　朋友重叠率指标问题示例 2

另外，朋友覆盖率指标在计算一条边的强度时没有考虑到相邻的其他边强度。在图 4-6 和图 4-7 中，按照朋友重叠率指标计算出的边权值相同，然而由于边权值不同，在图 4-7 中 A、B 与他们的共同朋友 C 与 D 关系强度远大于图 4-6 的情况，然而，按照常理判断，显然图 4-6 中的权值要大于图 4-7 中的权值。

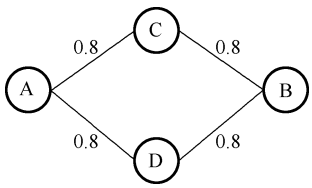


图 4-6　朋友重叠率指标问题示例 3

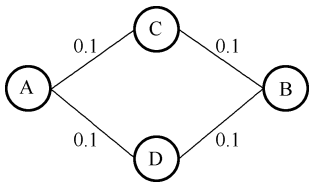


图 4-7　朋友重叠率指标问题示例 4

以上这些问题可以通过引入朋友间的聚类系数来解决。

在社交网络中，朋友的朋友很可能彼此也是朋友，这种属性称为网络的聚类特性。一般地，假设网络中的一个节点 i 有 k_i 条边与其他节点相连，这 k_i 个节点之间实际存在的边数 E_i 和总的可能边数 $k_i(k_i-1)/2$ 之比定义为节点 i 的聚类系数 C_i，即 $C_i=(2E_i/k_i(k_i-1))$。

整个网络的聚类系数 C 就是所有节点 i 的聚类系数 C_i 的平均值。大规模网络大多都具有明显的聚类效应。在社交网络中，朋友的朋友同时也是朋友的概率会随着网络规模的增大而趋于某个非零常数，这符合社会关系网络中“物以类聚，人以群分”的特性。

设 U_i 为节点 i 的邻居集合，U_j 为节点 j 的邻居集合，节点间朋友聚类系数计算公式定义如下：

$$w_{ij}=\frac{E_{ij}}{k_i\times k_j} \tag{4-17}$$

式中，E_{ij} 为 U_i 和 U_j 之间存在的边的加权和，k_i 为节点 i 的邻居数，k_j 为节点 j 的邻居数，w_{ij} 为两端分别为节点 i 和节点 j 的边的权值。

通过对两节点间聚类系数的计算，可以知道两个节点朋友圈之间的相似性。公式（4-17）不但考虑了朋友的重叠率，而且考虑了朋友集合间的相似度。按照该方法重新计算图 4-4 和图 4-5 中的例子，得出相同结果 $w_{ij}=1/1=1$，说明该方法可以避免由于节点本身朋友数量相差悬殊而产生的连接强度评价误差。

这样，按照用户关系强度模型［参见式（4-13）～式（4-15）］计算节点连接强度时，引入节点间朋友聚类系数，在相似度 $s=[s^{(a_1b_1)}s^{(a_2b_2)}\cdots s^{(a_Nb_N)}]^{\mathrm{T}}$ 中将 S_1 统一指定为节点间朋友聚类系数值，以减少连接强度计算误差。根据节点连接强度计算结果排名，找出弱连接。

4.4.2 信息传播模型

信息传播范围可以归结为影响力最大化问题，影响力最大化的关键问题是在网络中找出最有影响力的 k 个节点。这样，将社区影响力最大化问题变为选择最好的 k 个节点初始激活，目的是在影响力最大化过程的最终阶段使得社区覆盖最大化。信息传播过程可以从一个初始不活跃节点开始，随着时间的推移，某个节点的邻居节点中有越来越多的节点变为活跃；在某个时间点上，可能使该节点变为活跃。当一个节点在时间步 t 首先变为活跃时，可认为它具有感染力，拥有影响每个不活跃邻居节点的一次机会。一次成功的激活尝试将使其邻居在下一个时间步 t+1 变为活跃节点。如果某个节点的多个邻居节点在时间步 t 变为活跃，则这些活跃的邻居节点按任意顺序尝试激活他们的邻居节点，但所有的这些尝试都发生在时间步 t。一个活跃节点 u 对其所有邻居节点尝试激活后，仍保持活跃，但已不具备感染力了。当不存在具有感染力的节点时，信息传播过程结束。

根据以上的信息传播过程，可以在独立级联模型的基础上抽象出一个二元信息传播模型，通过将连接强度引入信息传播过程来描述社交网络的信息传播机制。在模型中引入参数 α 和 β，其中 α 为导航特征，它决定了如何选择邻居节点来再发布信息；$\beta \in [0,1]$ 为信息的强度，用于描述信息的重要或者有趣程度。

模型定义如下：

（1）假设有信息 I，设定 V 中所有节点状态为 σ_0，节点状态 σ_0 表示节点对信息 I 未知，否则变为状态 σ_1。

（2）从网络中随机选择一个种子节点 i，节点 i 的度数为 k_i，设置节点 i 状态为 σ_1，在时刻 $T=0$ 用强度 β 发布信息 I。

（3）增加一个时间单元 $T=T+1$，设置节点 i 的所有邻居节点状态为 σ_1，将节点 i 加入已发布信息 I 的节点集合 P，$P=P\bigcup\{i\}$。

（4）计算下一轮将要发布信息的节点数，即：

$$R_i = k_i\beta \tag{4-18}$$

（5）按概率 p_{ij} 选择节点 i 的一个邻居节点 j，即：

$$p_{ij} = \frac{w_{ij}^{\alpha}}{\sum_{m=1}^{k_i} w_{im}^{\alpha}} \tag{4-19}$$

如果 $j \notin P$，则将节点 j 加入下一轮将发布信息 I 的节点集合 W 中，$W=W\bigcup\{j\}$。重复步骤（5）进行 R_i 轮。

（6）对于节点集合 W 中的每个节点，递归执行步骤（3）～步骤（5），直到集合 W 为空或者集合 V 中所有节点收到信息 I。

从公式（4-18）可以看出，从节点 i 的邻居中选择再发布节点的数量取决于 k_i 和 β 的值。这种现象与实际情况一致：用户的朋友越多，该用户吸引来访问和再发布信息的用户也就越多；该条信息越有趣或越重要，则获得再发布的机会就越大。在公式（4-19）中使用参数 α 来表示信息传播中的连接强度，不同的 α 值代表下一轮再发布信息会选择不同路径进行。事实上，当 $\alpha=-1$ 时，信息再发布选择弱连接作为传播路径。当 $\alpha=0$ 时，传播路径随机选择。当 $\alpha=1$ 时，传播路径优先选择强连接。

4.4.3　模型验证

下面通过实验数据对社交网络的连接强度模型和信息传播模型进行测试和验证，同时对社交网络的弱连接及其影响进行分析。

1．实验数据集

实验数据来源于 CDBLP 网、Arvix 网、Wiki 网和 Enron 网等社交网站。

CDBLP 网是一个以作者为中心的中文学术作者合作网站，文献原始数据库中包括了中国计算机领域各个著名期刊历年文章作者的合作数据，其中作者的合作关系所构建的合作网络可以在一定程度上反映中国计算机领域的学者间合作情况。

Arvix 网与 CDBLP 网类似，为国外免费论文共享网站。

Wiki 网是一个由世界上众多志愿者合作完成的在线免费百科全书，在众多志愿者中有一小部分人是管理者。为了使一个普通用户能够变成管理者，Wiki 网使用了一种志愿者间相互投票决定的机制。该数据集已被众多的文章用来研究网络拓扑特性，比较有代表性。

Enron 网是一个电子邮件网络，其数据用来验证弱连接对通过电子邮件收发方式进行信息传播的影响程度。

2．最大连通子图

为了理解连接强度和网络结构之间的局部关系和全局关系，首先通过移除强连接或弱连接来观察网络的维持能力，然后通过测量最大连通分支的大小来评价移除连接对网络结构的影响。最大连通分支是指作为移除连接后剩余部分的节点通过相连的路径可以相互到达。由于对网络特性的相关分析只有在一个连通子图下才有意义，因此在实验之前需要抽取数据集中的最大连通子图作为最大连通分支。采用 UCINET 网络分析软件抽取最大子图作为最终研究数据，采用广度优先的搜索方法寻找最大连通子图，设 V 表示社交网络的节点集合，E 表示边的集合，则社交网络图记为 $G(V, E)$。如果两个节点之间有联系，则它们之间存在一条边；反之，节点间不存在边。对于网络图 $G(V, E)$ 中的一个节点，与边直接相连的节点称为节点的邻居节点。从图中的任意一个顶点出发，找出该顶点的等价类，然后再找出该顶点等价类中各元素的等价类，直到顶点等价类为空集，所得结果为最大连通子图。上述的 4 个网络数据集的最大连通子图如图 4-8～图 4-11 所示。

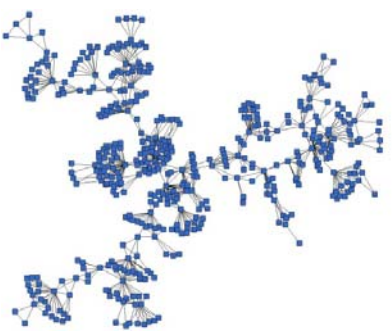


图 4-8　CDBLP 网数据集最大连通子图

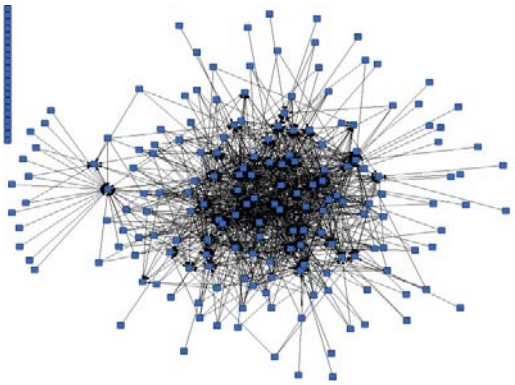


图 4-9　Arvix 网数据集最大连通子图

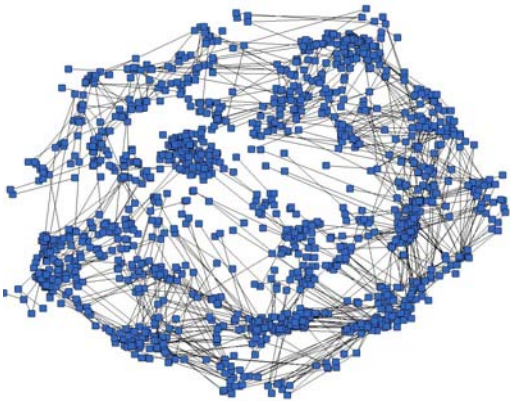


图 4-10　Wiki 网数据集最大连通子图

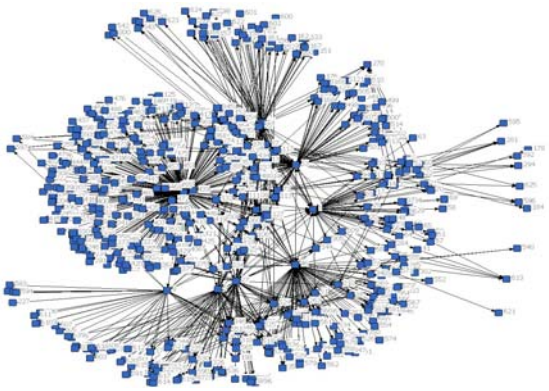


图 4-11　Enron 网数据集最大连通子图

数据集处理结果如表 4-3 所示。

表 4-3　数据集处理结果

网站名称	|V|	|E|	图示
CDBLP	572	3850	图 4-8
Arvix	7632	38671	图 4-9
Wiki	9311	143687	图 4-10
Enron	36692	367662	图 4-11

3．连接强度计算

首先，使用连接强度模型来计算各个数据集的连接强度及累积分布。在计算连接强度时需要提取各个数据集的相似度和交互活动两方面特征。

在计算相似度时，对于每一对用户 (a,b)，需要提取若干特征来表示用户配置文件之间的相似度。定义总体相似度为：$s=[s^{(a_1b_1)}s^{(a_2b_2)}\cdots s_i^{(a_Nb_N)}]^{\mathrm{T}}$。CDBLP 网、Arvix 网、Wiki 网和 Enron 网的相似度特征描述如表 4-4 至表 4-6 所示。

表 4-4　CDBLP 网、Arvix 网配置文件相似度特征

特征	描述
s_1	朋友聚类系数值
s_2	如果 *a* 和 *b* 曾在同一所学校则为 1，否则为 0
s_3	如果 *a* 和 *b* 在同一个公司则为 1，否则为 0
s_4	如果 *a* 和 *b* 在同一个城市则为 1，否则为 0
s_5	如果 *a* 和 *b* 在同一个行业则为 1，否则为 0
s_6	如果 *a* 和 *b* 在同一个研究领域则为 1，否则为 0
s_7	如果 *a* 和 *b* 之间的职称相同则为 1，否则为 0

表 4-5　Wiki 网配置文件相似度特征

特征	描述
s_1	朋友聚类系数值
s_2	如果 *a* 和 *b* 在同一个城市则为 1，否则为 0
s_3	如果 *a* 和 *b* 在同一个行业则为 1，否则为 0
s_4	如果 *a* 和 *b* 在同一个研究领域则为 1，否则为 0
s_5	如果 *a* 和 *b* 之间的职称相同则为 1，否则为 0

表 4-6　Enron 网配置文件相似度特征

特征	描述
s_1	朋友聚类系数值
s_2	如果 *a* 和 *b* 在同一个公司则为 1，否则为 0
s_3	如果 *a* 和 *b* 在同一个城市则为 1，否则为 0
s_4	如果 *a* 和 *b* 在同一个行业则为 1，否则为 0
s_5	如果 *a* 和 *b* 都在对方的地址簿中则为 1，否则为 0

对于用户交互活动，同样需要提取各种类型的用户交互特征来表示。表 4-7 显示了 CDBLP 网、Arvix 网、Wiki 网的交互特征，分别用 $m_1^{(ab)}$、$m_2^{(ab)}$、$m_3^{(ab)}$ 和 $m_4^{(ab)}$ 来表示。表 4-8 显示了 Enron 网的交互特征，分别用 $m_1^{(ab)}$、$m_2^{(ab)}$ 来表示。

表 4-7　CDBLP 网、Arvix 网、Wiki 网交互特征

特征	描述
m_1	如果 a 和 b 之间有连接则为 1，否则为 0
m_2	如果 a 和 b 之间有邮件活动则为 1，否则为 0
m_3	如果 a 和 b 之间有评论行为则为 1，否则为 0
m_4	如果 a 和 b 之间有合作则为 1，否则为 0

表 4-8　Enron 网交互特征

特征	描述
m_1	如果 a 和 b 之间有连接则为 1，否则为 0
m_2	如果 a 和 b 之间有邮件活动则为 1，否则为 0

在模型中，每种类型的交互活动中包括一个辅助变量，用来表示与用户 a 存在指定方式交互活动的总人数。例如 m_1 的辅助变量表示与节点 a 建立连接的总节点数。

各个数据集的连接强度计算结果及累积分布如图 4-12 所示，从图 4-12 可以看出，各个数据集的弱连接分布是不同的，在 CDBLP 网和 Arvix 网数据中连接强度小于 0.2 的部分占比更大一些。

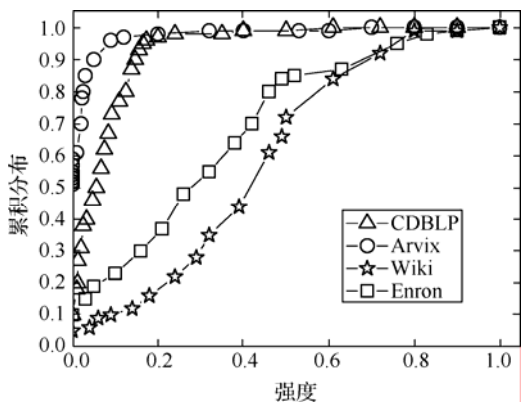


图 4-12　连接强度及累积分布

4．对网络结构影响分析

为了分析弱连接对网络结构的影响，需要定义一个动态指标[9]来监测网络结构的变化，即：

$$\overline{S}=\sum_{S<S_{\max}}\frac{nS^2}{N} \tag{4-20}$$

式中，n 为 S 个节点所连接的集群的数量，N 为网络规模的大小。

最大连通子图的相对值用 $R_{GC}(f)$ 来表示，即：

$$R_{GC}(f)=N_{GC}(f)/N_{GC}(f=0) \tag{4-21}$$

式中，f 为节点之间关系强度，N 为网络的大小。

在网络结构影响实验中，首先将节点按照连接强度排序，分别按照从强到弱和从弱到强的顺序从网络中移除节点，并按照公式（4-20）计算动态指标 $\overline{S}$，按照公式（4-21）计算最大连通子图相对值 $R_{GC}(f)$。

CDBLP 网和 Arvix 网数据集移除连接后的稳定性实验结果如图 4-13～图 4-16 所示。

图 4-13 中参数 f 代表按照朋友重叠率移除连接的部分，小三角形曲线表示从弱到强的连接移除操作过程，小矩形曲线表示从强到弱的连接移除操作过程。从图 4-13 中 $R_{GC}(f)$ 的变化趋势可以看出，当从弱到强移除连接时，会导致网络的突然崩溃；而从强到弱移除连接时，网络只会逐渐缩小而不会崩溃。当弱连接首先被移除时，在某个值（大约 0.55）处，三角形曲线变为 0。

图 4-14 与图 4-13 的 $R_{GC}(f)$ 变化趋势类似，当从弱到强移除连接时，同样会导致网络的突然崩溃；而从强到弱移除连接时，网络只会逐渐缩小而不会崩溃。当弱连接首先被移除时，在某个 f 值（大约 0.76）处，三角形曲线变为 0。

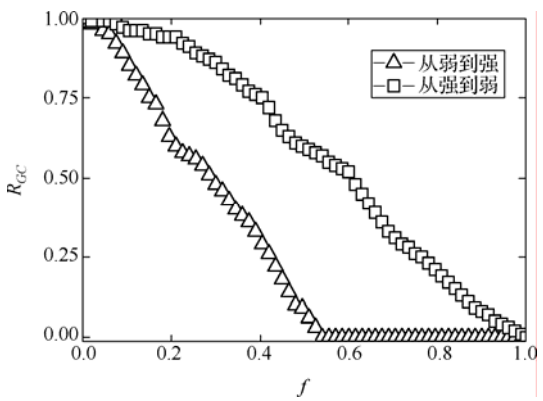


图 4-13　CDBLP 网的 R_{GC}

从图 4-15 可以看出，当从弱到强移除连接时，曲线上有一个明显的尖峰；而当从强到弱移除连接时，并没有出现尖峰。在 $N\to\infty$ 时，$\overline{S}$ 值接近临界阈值 f_c 时发散，f_c 即为移除连接的相变点。根据渗透理论，相变的存在表示网络崩溃。对于 CDBLP 网，相变点出现在 $f_c=0.548$ 处。

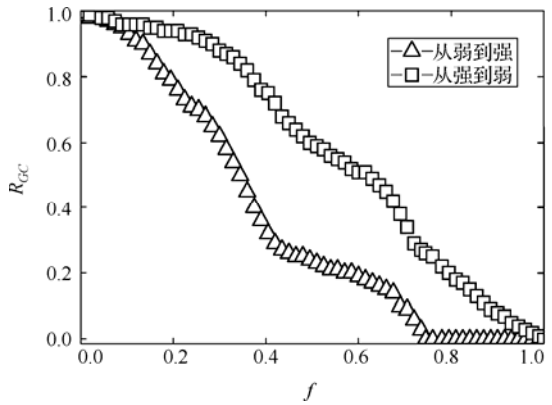


图 4-14　Arvix 网的 R_{GC}

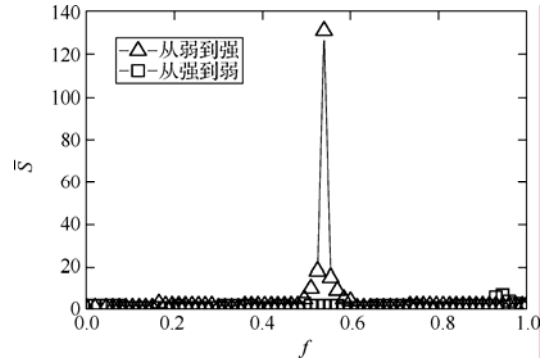


图 4-15　CDBLP 网的 $\bar{S}$

从图 4-16 可以看出，Arvix 网的相变点出现在 $f_c = 0.745$ 处。事实上，与强连接相比，弱连接更多存在于网络社区之间，从图 4-8 和图 4-9 给出的最大连通子图可以直观地看到，弱连接更像一些连接孤立社区之间的桥梁。

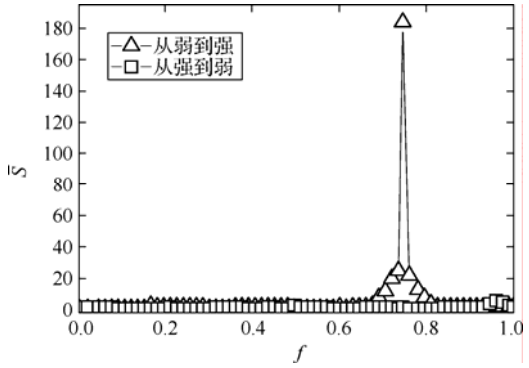


图 4-16　Arvix 网的 $\bar{S}$

Wiki 网和 Enron 网数据集移除连接后的稳定性实验结果如图 4-17～图 4-20 所示。图 4-17 和图 4-18 中参数 f 代表按照朋友重叠率移除连接的部分，小三角形曲线表示从弱到强的连接移除操作过程，小矩形曲线表示从强到弱的连接移除操作过程。与图 4-13 和图 4-14 比较，图 4-17 和图 4-18 中无论以何种顺序移除连接，最大连通子图的相对值 $R_{GC}(f)$ 的变化趋势都很稳定，是逐渐降低的。

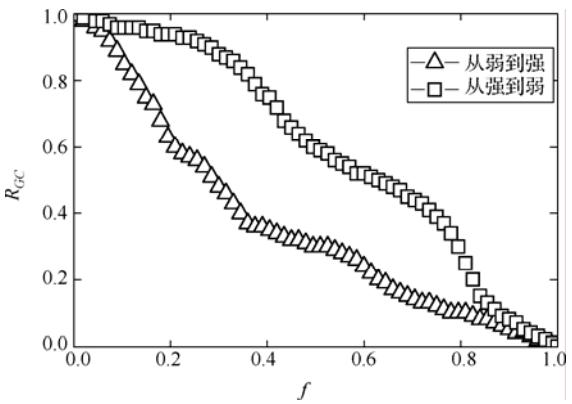


图 4-17　Wiki 网的 R_{GC}

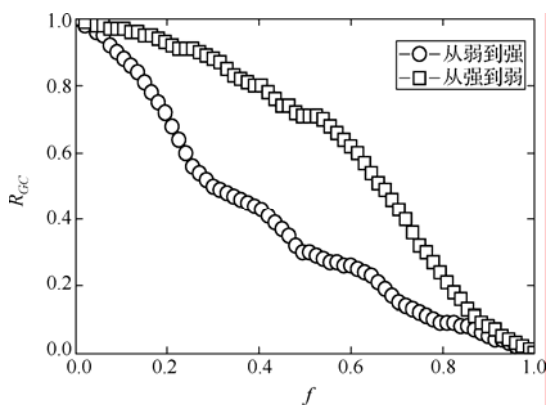


图 4-18　Enron 网的 R_{GC}

图 4-19 和图 4-20 中也没有明显的相变点，曲线上没有出现尖峰，这说明在移除连接过程中网络一直在逐渐缩小，并没有发生崩溃的现象，与图 4-15 和图 4-16 的结果完全不同。

实验结果表明，在社交网络中，从弱连接开始由弱到强的连接移除操作，对 CDBLP 网和 Avix 网的网络结构影响非常明显，而对 Wiki 网和 Enron 网影响并不大。因此，可以合理地推断，弱连接对社交网络结构的影响与网络的具体形式有关。

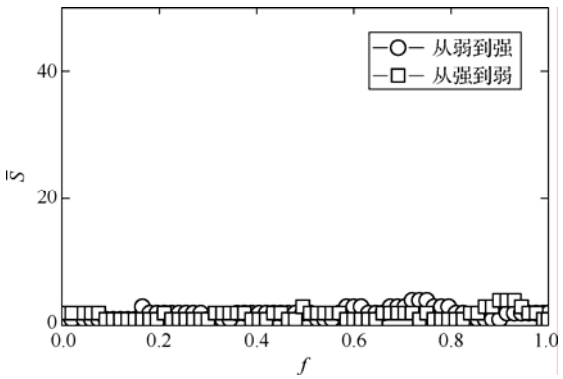


图 4-19　Wiki 网的 $\bar{S}$

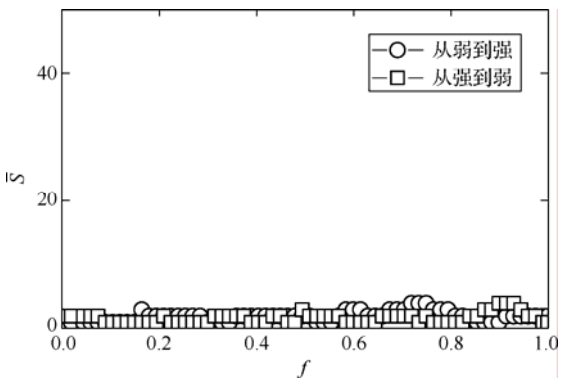


图 4-20　Enron 网的 $\bar{S}$

CDBLP 网和 Arvix 网属于基于朋友关系的实体关系网络，与传统的人际传播网络结构类似，其传播网络结构如图 4-21 所示。在这种传播网络结构中，信息传播具有对象选择性和节点高聚集性等特点，传播对象的选择性是指行动者不会随意将信息传递给与他发生接触的任何人，而是有选择地传播给他认识的其他行动者。对象选择性限制了传统人际传播网络中每个行动者能够直接联系的其他行动者的数量，即每个节点对外连接的边的数量。节点高聚集性造成原始信息的传播范围和传播路径相当有限，当去除聚集之间的弱连接后，信息很难传播到整个网络。

Wiki 网和 Enron 网属于互联网信息传播网络，其传播网络结构如图 4-22 所示。互联网信息传播网络有着类似的传播结构，网络采取网状结构进行信息传播和交流，强连接群体之间存在多条可达的连接，去除部分弱连接并不会影响信息的传递。

从以上分析可以看出，两类信息传播网络之间存在着如下差别：

（1）人际传播网络中存在着明显的社区特性，网络由多个联系紧密的社区组成。然而，在现代信息传播网络中并没有此类特性。

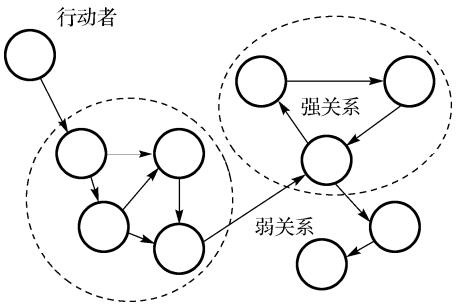


图 4-21　传统人际传播网络结构

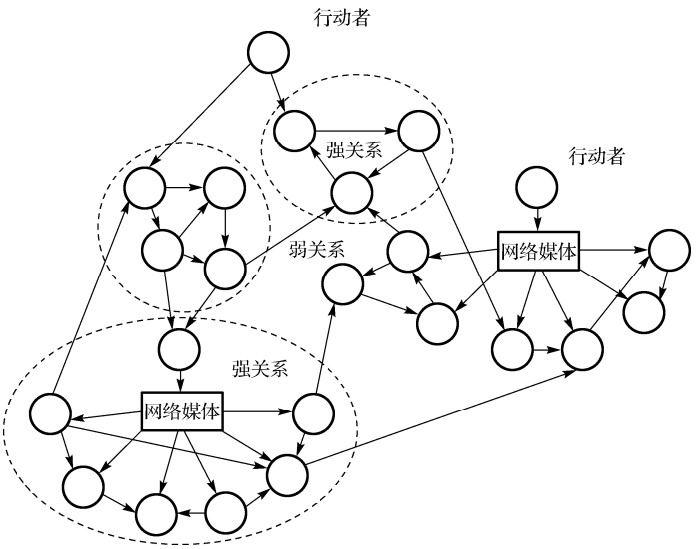


图 4-22　互联网信息传播网络结构

（2）互联网信息传播网络中节点的度数差异很大，有很多节点的度数在 200 以上，同时又有很多的节点度数仅为 1，大多数的节点度数相对稳定。通过统计分析发现，合作关系网络中的度分布基本服从正态分布。

5．对信息传播影响分析

在社交网络上传播的信息包括博客、图片、消息、评论、多媒体文件、状态说明等，由于社交网络上隐私控制等原因，其信息传播机制与传统方式有所不同。社交网络上的传播过程可以概括如下：

（1）用户 i 发布信息 I，可以是一张图片，也可以是一篇博客。

（2）用户 i 的朋友们通过某种方式知晓了消息 I，可以是自己访问 i 的配置页面，也可以是社交网站直接推送的。

（3）用户 i 的朋友，可能有一个、多个或者一个也没有，认为信息 I 很重要或者很有趣，所以评论、引用或者转发了 I，这种行为叫做信息再发布。

（4）随着信息再发布的人代替用户 i 的位置，上述的步骤会再次重复进行。

社交网络一个显而易见的关键特征是网络站点在积极推送信息而只有一部分用户会再发布这条信息。这个特征在微博网站上表现最为明显。例如，在新浪微博上，用户发布的所有微博内容都会立即推送给他的所有关注者的终端上，包括 PC、掌上电脑和移动设备。如果有关注者喜欢这条信息，他可以对信息再发布。然而在大规模网络上信息的传播轨迹是非常难收集的，因此必须通过建立某种模型来描述信息传播机制并模拟传播过程。4.4.2 节中的信息传播模型就是用来描述社交网络信息传播机制并模拟传播过程的。

定义所有状态为 σ_1 的节点集合为 C，表示信息 I 的传播范围。参考 Flickr 网站的用户信息再发布比率只有 1～2%[10]，设模型参数 $\beta=0.01$。图 4-23 显示了 CDBLP 网数据集信息传播实验结果，图 4-24 显示了 Wiki 网数据集信息传播实验结果。从图 4-23 和图 4-24 可以看出，当 $\alpha=0$ 时，集合 C 达到最大值。这就是说，与弱连接或强连接相比，随机选择信息再发布节点能够使信息传播速度更快、范围更广。

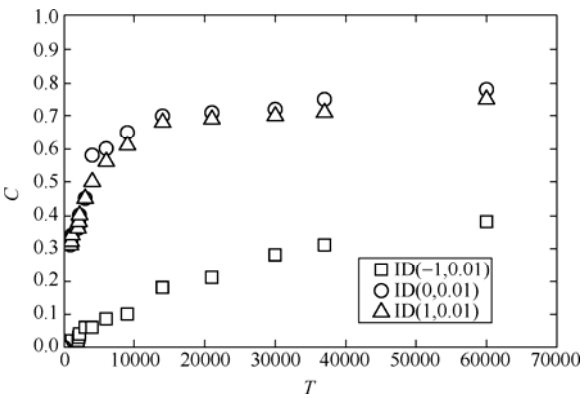


图 4-23　CDBLP 网数据集信息传播实验结果

下面结合信息传播细节来分析这个实验结果。由公式（4-16）可以得出：

$$1/w_{ij}=(k_i-2)/c_{ij}+k_j/c_{ij}-1$$

假设 k_j 增加时 c_{ij} 也成比例增加，即 k_j/c_{ij} 为常数，给定一个节点 i 和它的邻居节点 j，则有：

$$k_j\uparrow\Rightarrow c_{ij}\uparrow\Rightarrow 1/w_{ij}\downarrow\Rightarrow w_{ij}\uparrow$$

反之亦然。这意味着节点 i 的一个强连接邻居节点具有更高的度值，因此在实验中使用 α 值来代表所选择的再发布节点类型，不同的 α 值表示被选节点的度值也不同。例如，当

$\alpha=-1$时，弱连接邻居节点被选中作为再发布节点，而这些节点的度值很低，节点的度值越低，则再发布节点可选的邻居节点就越少，这最终会降低再发布节点的总数，从而阻碍信息在网络中的传播。当$\alpha=1$时，优先选择强连接，虽然选择的再发布节点的度值高，但由于节点选择限制在本地社区范围，信息在网络的传播范围也是有限的。

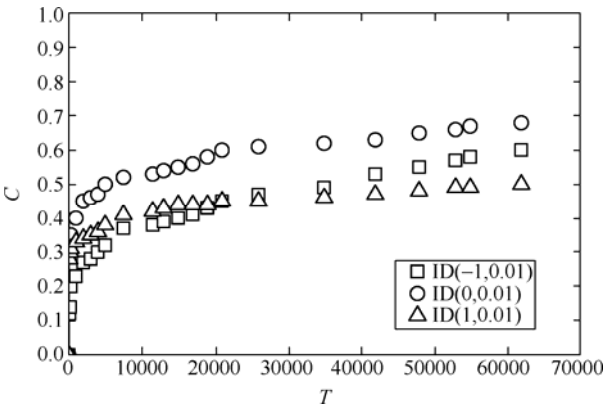


图 4-24　Wiki 网数据集信息传播实验结果

为了验证以上的分析，定义函数 f_{inf} 来表示随着时间进行信息发布的节点总数，图 4-25 和图 4-26 显示了信息传播过程中 f_{inf} 的变化情况。

在图 4-25 和图 4-26 中，当$\alpha=-1$时，f_{inf} 增大缓慢，不同 α 值下曲线变化趋势与图 4-23 和图 4-24 中的 C 曲线变化趋势相似。当$\alpha=1$时，随着传播过程的进行，f_{inf} 增大越来越慢，这是因为选择的再发布节点被限制在本地社区内部，从而很难发现新的节点能够让信息传播到远处。

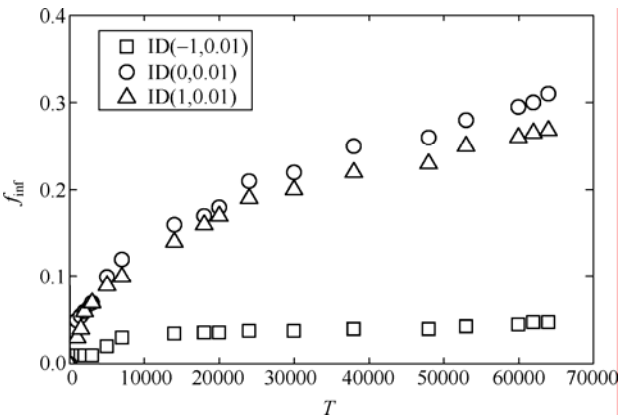


图 4-25　CDBLP 网数据集信息传播过程 f_{inf} 变化

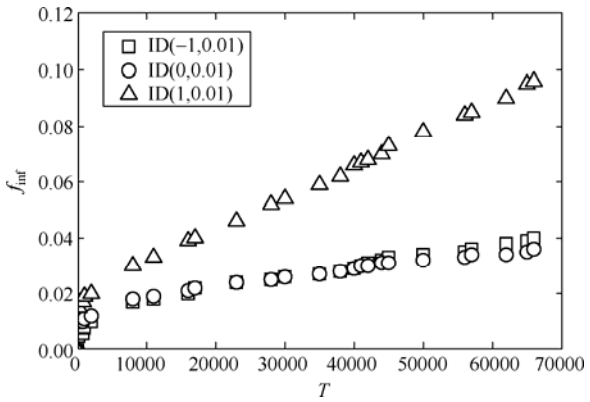


图 4-26　Wiki 网数据集信息传播过程 f_{inf} 变化

定义函数 f_{loc} 来表示参与信息发布的节点距离源节点的跳步数，图 4-27 和图 4-28 显示了信息传播过程中 f_{loc} 的变化情况。在图 4-27 和图 4-28 中，与 $\alpha=0$ 和 $\alpha=1$ 情况相比，$\alpha=-1$ 时的 f_{loc} 减少速度要快得多。这意味着当信息传播远离源节点时，如果使用弱连接来选择再发布节点，可选节点数会急速减少，这与前面的分析结果相一致。

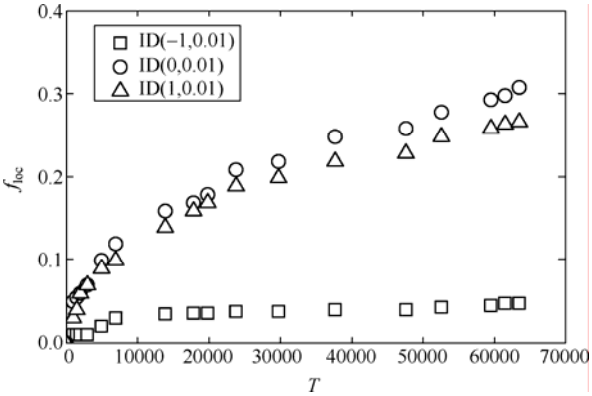


图 4-27　CDBLP 网数据集信息传播过程 f_{loc} 的变化

下面分析参数 β 的效果，图 4-29 和图 4-30 显示了不同 β 值的实验结果。从图 4-29 和图 4-30 可以看出，无论 β 如何取值，随机选择再发布节点总是信息传播最快的方法，当 β 值增大时，几种节点选择方法之间的差距会缩小。对于所有的 α 值，随着 β 值的增大，C 值都在增大，这表明信息的强度越大，就有越多的节点被吸引进来参加信息再发布，而信息在网络上的传播范围也就更广。

从以上的实验结果可以得出结论，优先选择弱连接作为信息再发布的路径并不能加快信息传播速度。但是，在特定形式的社交网络上，弱连接对信息传播覆盖范围的影响是很大的。

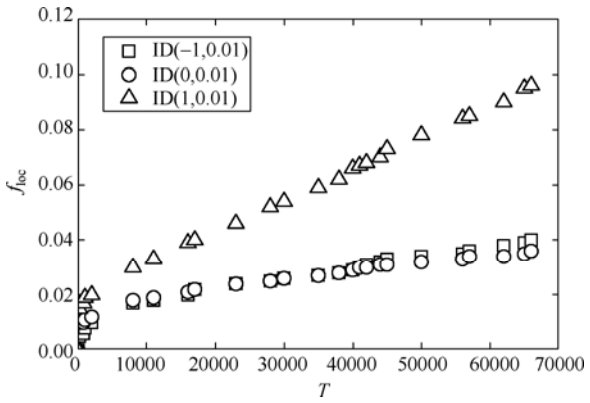


图 4-28　Wiki 网数据集信息传播过程 f_{loc} 的变化

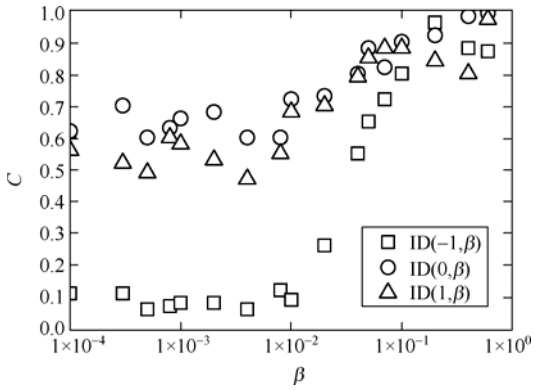


图 4-29　CDBLP 网数据集 β 按对数增长时 C 的变化

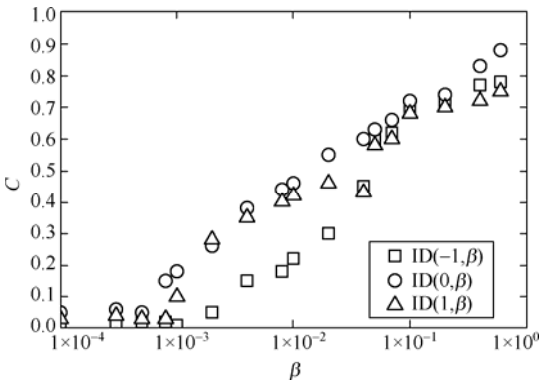


图 4-30　Wiki 网数据集 β 按对数增长时 C 的变化

4.5　社交网络用户关系预测

社交网络显著扩大了社交圈子，用户可以通过社交网络关注任意一个精英人物，例如政治家、明星、演员、运动员，或者自己的亲密朋友。当用户关注一个人并建立联系后，这个人会不会反过来也关注该用户呢？举例来说，一个微博用户关注了一个名人之后，这个名人会反过来也关注这个用户吗？一般情况下名人是不会关注普通用户的，但也不排除偶然情况的发生。有一些微博顶级用户有着几十万甚至上百万的关注者，而他们对自己的每一个关注者都进行了反关注。当其他人在查找关注列表进行手工添加新关注的时候，一些人却在使用工具进行自动的反向关注操作。掌握这些关系的建立方式有助于很多网络应用的研究与开发，例如朋友推荐、社区发现、产品推广等。

下面主要研究社交网络中用户双向关系预测问题。

4.5.1　用户关系特征

在社会科学中，两个个体之间的关系分为两种类型：单向关系和双向关系。单向关系称为准社会交往，双向关系称为互惠交往。一般情况下，名人和拥趸之间的关系是一种单向关系，而亲密朋友之间的关系是双向关系。新浪微博和人人网就是这两种关系类型的典型代表。

下面以新浪微博数据为基础，分析一个单向关系是如何发展成为一个双向关系的。当一个用户关注一个特定的用户（如某个名人）时，这个名人是否会反向关注该用户，即当用户给一个人发送了朋友请求之后，对方有多大的可能性会确认该用户的朋友请求。这个问题同样存在于人人网、QQ 网等社交网站上。

1．问题定义

这里结合新浪微博的实际情况来定义问题和框架。微博网络可以被模型化为一个有向图 $G=\{V,E\}$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 是用户集合，$E\subseteq V\times V$ 是用户之间有向连接的集合。每个有向连接 $e_{ij}=(v_i,v_j)\in E$ 代表用户 v_i 关注用户 v_j。

随着连接的增加和删除，微博网络在动态变化。通过对新浪微博数据的初步统计，发现用户增加连接的行为要比删除现有连接的行为频繁得多，大约 97%的连接变化是在增加新连接。因此增加的新连接构成了新浪微博的网络结构，一个新连接代表在微博上一个用户对另一个用户进行了关注或反向关注。

关注行为定义如下：假设在时刻 t，用户 v_i 建立了一个到 v_j 的连接，而 v_j 之前没有到 v_i 的连接，那么称 v_i 对 v_j 执行了一个新关注行为。当用户 v_i 在时刻 t 建立了一个到 v_j 的连

接，而在时刻 t 之前 v_j 已经有到 v_i 的连接，那么称 v_i 对 v_j 执行了一个反向关注行为。

新关注和反向关注行为分别对应社会学中的单向关系和双向关系，这里主要关注反向关注行为的形成过程。为简单起见，用 $y_{ij}^t=1$ 表示用户 v_i 在时刻 t 反向关注用户 v_j，$y_{ij}^t=0$ 表示用户 v_i 没有进行反向关注行为。这样，将关注行为预测问题描述如下：$<1,\cdots,t>$ 表示有特定时间粒度的时间戳序列，时间粒度可以是一天或者一周，微博网络用时间描述为 $\{G^t=(V^t,E^t,Y^t)\}$，其中 Y^t 表示在时刻 t 反向关注行为的集合，则有预测函数 $f:(\{G^1,\cdots,G^t\})\rightarrow Y^{(t+1)}$，可以用该函数推测在时刻（$t$+1）的反向关注行为。

这里对问题的描述与已有的链接预测和社会行为预测问题完全不同。首先，由于微博网络是随着时间进化的，在时刻 t 收集整个网络的数据是不可行的，所以设计一种方法能够同时兼顾到那些未能采集到的数据是非常重要的。其次，形成反向关注关系的基本特征尚不明确，需要将社会学理论和统计学等特征综合在一个统一的模型中，以便于更好地预测反向关注关系。

2．特征因素分析

特征因素分析实验使用新浪微博数据集。数据集抓取于 2011 年 5 月至 7 月。随机选取了新浪微博 3430 个种子节点，并抓取所有种子节点关注者的信息，其中包括微博、用户资料、用户标签、关注行为等。在微博用户中，存在一部分很少发布微博的不活跃用户，这些不活跃用户几乎不参与微博的互动，因此有必要将这些用户从数据集中排除。

定义 θ 为用户活跃指数，$\theta=T/R$，其中 T 表示发布微博的数量，R 表示用户注册时间。T 和 R 的信息可以从用户资料中获取，设定阈值 $\theta=2$，即平均每天发布大于 2 条微博的用户被认为是活跃用户，否则为不活跃用户。通过数据预处理，最终的数据集包括了 3430 个种子用户，171769 活跃关注用户，702789 活跃边，其中 185327 条活跃边有转发记录。

在实验中，把每个用户的地理位置特征也考虑进去。利用用户资料填写的省份与城市来定义两个用户地理距离的远近，可分为三类：

（1）近距，表示两个用户处在相同城市；

（2）中距，表示两个用户处在不同的城市但是相同省份；

（3）远距，表示两个用户处在不同省份。

不同的特征对建立反向关注有不同的影响，这些特征影响因素如下：

（1）地理距离：用户之间处于同一个地区时相互关注的概率是否更高；

（2）同质性：用户之间相似度越高是否能够越相互关注；

（3）隐式网络：微博的关注网络是否与其隐藏的转发网络和评论网络相对应；

（4）结构平衡：微博的双向关系网络在多大程度上满足结构平衡理论。

1）地理距离

地理距离与两个用户建立一个双向关系的概率之间对应关系如图 4-31 所示，图 4-31（a）中的 X 轴值表示不同省份，Y 轴值表示一个省份内的用户反向关注另一个用户的概率，图 4-31（a）中显示了用户对其他来自相同或不同省份的用户进行反向关注的可能性，很显然各个柱状体差别不大，这表明地理距离并不能成为阻挡人们建立双向关系的阻碍特征。图 4-31（b）中的 X 轴表示同城、同省不同城市和不同省，Y 轴表示用户相互关注的数量，从图 4-31（b）中可以看出，同城用户相互关注的数量远远大于地理距离较远的用户对，这表明微博网络仍然具有本地性。

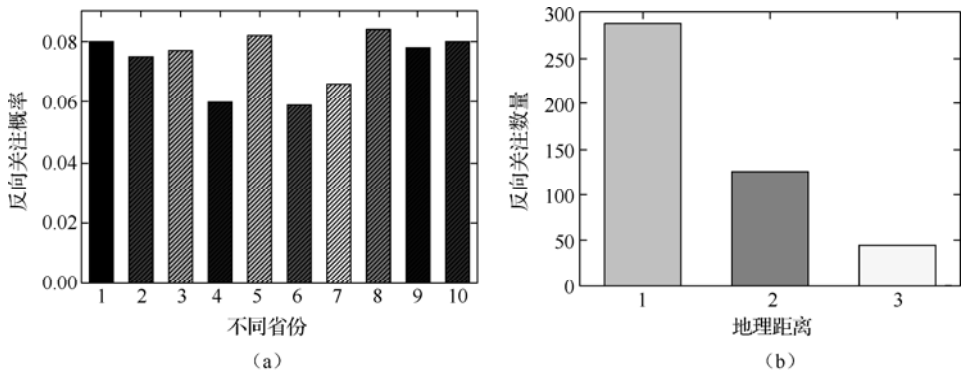


图 4-31　地理距离相关性

2）同质性

同质性是指具有相似性的用户之间倾向于相互建立联系的特征，这些特征可以是年龄相近或者社会地位相近等。这里主要研究微博网络上的两种同质性：连接同质性和状态同质性。

对于连接同质性，通过检测用户之间是否有共同的连接来确定他们彼此之间是否有建立联系的趋势，共同的连接是指两个用户具有共同的关注者或者共同的被关注者。连接同质性特征分为 4 种情况：共同的邻居数、共同的双向连接数、共同的关注者数量和共同的被关注者数量。连接同质性如图 4-32 所示，图中的 X 轴表示共同邻居数，Y 轴表示反向关注的概率。从图 4-32 可以明显看出，具有共同邻居的两个用户之间彼此相互反向关注的概率要远大于没有反向关注的用户。当双向关注的共同邻居数增加到 3 时，两个用户彼此反向关注的概率也增加到三倍。当共同邻居数增加到 10 时，概率增加效果更为明显。但是这种情况只适用于双向关注而不适用于单向关系。

对于状态同质性，通过检测用户之间是否有相似的社会状态来确定他们彼此之间是否更有可能建立联系。状态同质性特征分为三种情况：是否都是精英用户、是否一个精英用

户一个普通用户、是否都是普通用户。这里通过三种不同的算法将用户分成精英用户和普通用户两部分，三种算法是：PageRank[11]、度计算和 (α,β) 算法[12]。

（1）PageRank 算法是一种用来估计网络中节点重要性的算法，使用该算法可以估计每个用户在网络结构中的重要性，按照评分将排名前 1%的用户认定为精英用户，其他用户归为普通用户部分。

（2）度计算法是一种用来计算用户节点度值的方法，使用该方法将入度最高的前 1%用户认定为精英用户，其他用户归为普通用户部分。

（3）(α,β) 算法是一种用来发现社交网络核心成员的算法，使用该算法可以将核心社区的大小设置为 200，然后将选择出的社区核心用户作为精英用户，剩余其他用户为普通用户。

使用三种算法计算的状态同质性如图 4-33 所示，图 4-33 中的 X 轴表示不同算法和不同类别用户，Y 轴表示反向关注的概率。从图 4-33 可以明显看出，尽管三种算法的统计结果不同，但是它们的共同之处是精英用户之间相互关注的趋势最强。在 (α,β) 算法下，精英用户之间反向关注的可能性比普通用户之间要高出 5 倍。(α,β) 算法似乎更加符合对问题的设定，能够更好地区分精英用户和普通用户，其主要原因是该算法在网络结构中考虑了面向精英用户的社区结构。

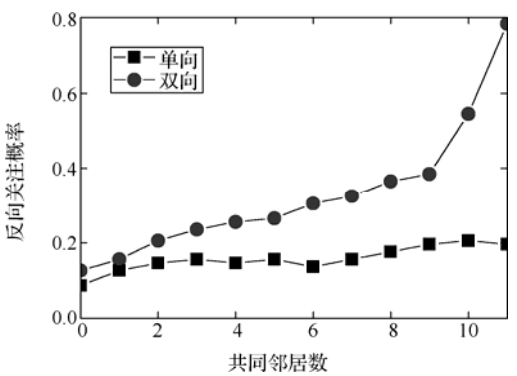


图 4-32　连接同质性

3）隐形网络

在微博上除了直观的关注连接网络外，还有一些可以用结构信息推导出来的隐形网络结构。例如：用户 A 在他的微博里提到了用户 B，A 使用@B 做了一个评论连接；用户 A 在微博中使用了 B 的微博，这就产生了一个转发连接。微博上这些隐形连接特征分为 4 种情况：为 A 对 B 的评论、A 对 B 的转发、B 对 A 的评论和 B 对 A 的转发。隐形网络相关性如图 4-34 所示，图中的 X 轴表示回复和转发，Y 轴表示反向关注的概率。图 4-34

显示出用户 A 和 B 彼此评论和转发对方微博时，他们之间反向关注的可能性也比较高。另外，转发一个人的微博比单纯评论更加有助于获得对方的反向关注。

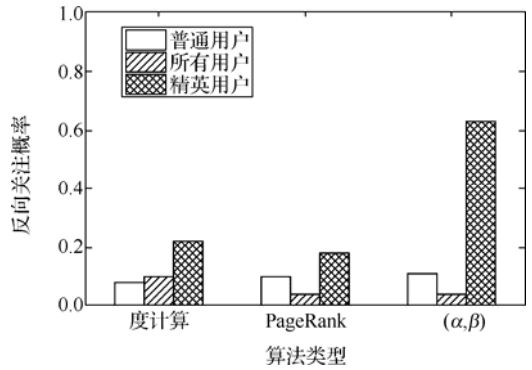


图 4-33　不同算法下的状态同质性

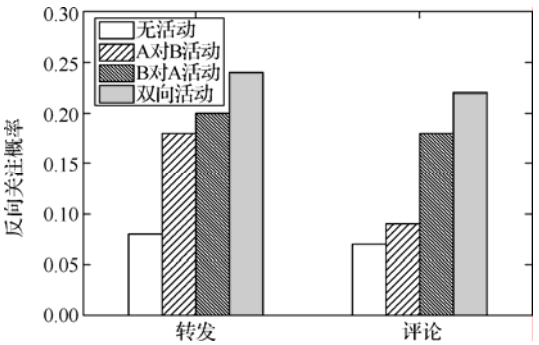


图 4-34　隐形网络相关性

4）结构平衡

结构平衡理论是一个基本的社会心理学理论[13]。每一个三用户组被称为三元组，结构平衡性是指三个用户相互之间都是朋友，或者只有其中一对用户是朋友。结构平衡特征分为 4 种情况，参见图 4-35。图 4-35 中的（a）和（b）是平衡结构，而（c）和（d）是不平衡结构。可以将朋友关系图形转化为双向关系或单向关系，然后运用结构平衡理论来检验微博网络上的关系是否满足结构平衡性。图 4-36 中显示了大约有 88%的双向连接用户通过一个平衡结构相连，而单向关系中的连接结构却是不平衡的，这是由于两个用户都关注同一个精英用户但是他们彼此之间却并不相识造成的，这种不平衡三元组结构如图 4-35（c）所示。

根据对以上 4 种特征因素的统计分析，得出如下结论：

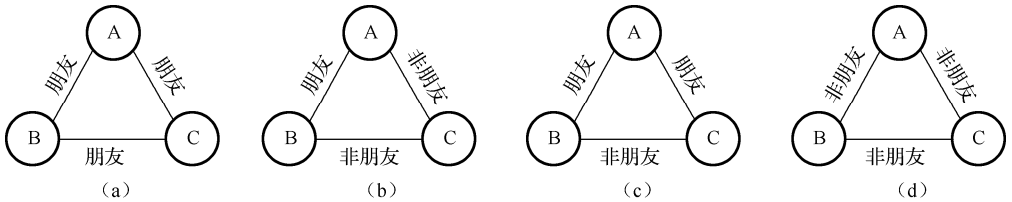


图 4-35　结构平衡理论图解

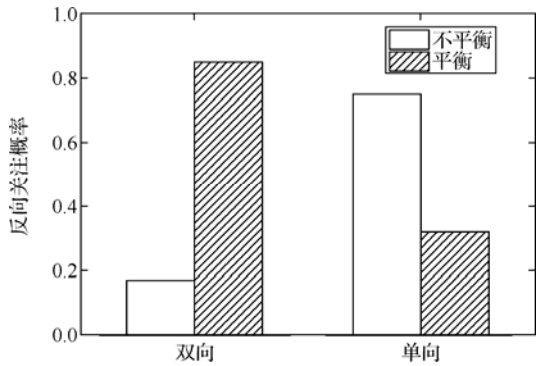


图 4-36　结构平衡相关性

（1）地理距离对于用户之间建立双向关系的数量具有显著影响，但是对用户建立反向关注的可能性影响很小，具有共同朋友的用户之间更加倾向于彼此相互关注。

（2）精英用户比普通用户具有更强的相互关注的趋势。

（3）微博转发和微博评论隐形网络与双向关系的形成有很强的相关性。

（4）微博上双向关系网络结构是平衡的，而单向关系网络结构是不平衡的。

4.5.2　预测模型

为了准确地预测双向关系，采用一种三元特征模型，将所有信息都包含在一个单一实体里。将边表示为 e_i，边两端的用户表示为 v_i^s 和 v_j^s，假设时刻 t 用户 v_i^s 关注 v_j^s，需要预测时刻（t+1）用户 v_j^s 是否反向关注 v_i^s。这里需要定义边的一些属性，用 x_i 来表示，属性矩阵 $X=|E|\times d$ 描述每条边的具体属性，d 表示属性的数量。例如，微博上的一个属性可以定义为两个用户是否处于同一个省份。矩阵 X 的元素 x_{ij} 代表边 e_i 的第 j 个属性值。三元特征模型来自于社会学理论，将结构平衡理论中的三元组概念包含在模型之中。

图 4-37 显示了三元特征模型的图形结构，图 4-37 分为左侧和右侧两个部分，左侧部分显示的是 5 个用户在时刻 t 的关注关系网络，空心箭头代表新增加的关注行为，实心箭头代表时刻 t 之前的关注行为，符号“/ ”代表在时刻 t 用户 v_i^u 没有关注用户 v_i^s。图 4-37

中右侧部分是从左侧输入网络得到的特征模型，右下框中的椭圆代表用户关系 (v_i^u, v_i^s)，右上框中边的标签 y_i 是一个隐藏变量。当 $y_i=1$ 时，表示 v_i^u 进行了一个反向关注行为，当 $y_i=0$ 时，则没有关注。y_i 正是需要进行预测的变量。特征 $h(\cdot)$ 表示一个定义在三元组之上的平衡因子函数，$f(v_i^s, v_i^u, y_i)$ 代表边 e_i 辅助信息的一个特征，可以表示为 $f(x_i, y_i)$。

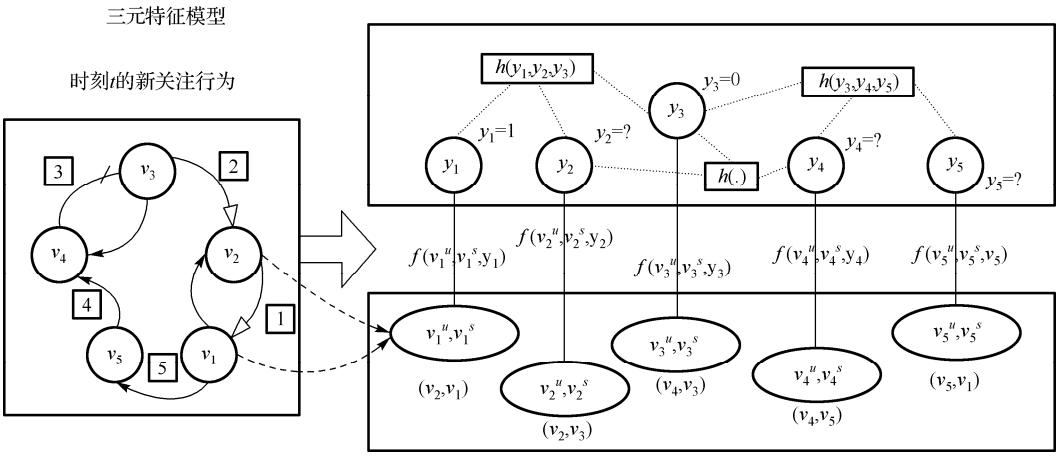


图 4-37　三元特征模型图形表示

时刻 t 的网络为 $G^t=(V^t, E^t, X^t)$，有已知变量 $y=1$ 和 $y=0$，还有一些未知变量 y 的值需要推测。为简化表示去除上标 t。计算网络的先验概率 $P(Y|X,G)$，根据 Bayes 定理，则有：

$$P(Y|X,G)=\frac{P(X,G/Y)|P(Y)}{P(X,G)}\propto P(X|Y)\cdot P(Y|G) \tag{4-22}$$

式中，$P(Y|G)$ 为网络结构标签的概率。

$P(X|Y)$ 表示与每条边标签 Y 相联系的属性 X 的生成概率。假设由每条边标签给出的属性的生成概率是条件独立的，则有：

$$P(Y|X,G)\propto P(Y|G)\prod_i P(x_i|y_i) \tag{4-23}$$

式中，$P(x_i|y_i)$ 为由标签 y_i 给出的属性 x_i 的生成概率。

下面实例化概率 $P(Y|G)$ 和 $P(x_i|y_i)$，使用马尔科夫随机场建模，通过 Hammersley-Clifford 定理将两个概率实例化为：

$$P(x_i|y_i)=\frac{1}{Z_1}\exp\left\{\sum_{j=1}^{d}\alpha_j f_j(x_{ij}, y_i)\right\} \tag{4-24}$$

式中，Z_1 为标准化因子，x_{ij} 为每个与边 e_i 相对应的属性，α_j 为第 j 个属性的权重，$f_j(x_{ij},y_i)$ 为 x_{ij} 的特征函数。

$$P(Y|G)=\frac{1}{Z_2}\exp\left\{\sum_c\sum_k \mu_k h_k(Y_c)\right\} \tag{4-25}$$

式中，Z_2 为标准化因子，μ_k 为第 k 个相关特征函数的权重，$\{h_k(Y_c)\}_k$ 为网络中每一个三元组 Y_c 建立的一组相关特征函数。

根据公式（4-22）、（4-23）、（4-24）、（4-25），定义关注行为的对数似然目标函数 $O(\theta)=\log P_\theta(Y|X,G)$，即：

$$O(\theta)=\sum_{i=1}^{|E|}\sum_{j=1}^{d}\alpha_j f_j(x_{ij},y_i)+\sum_c\sum_k \mu_k h_k(Y_c)-\log Z \tag{4-26}$$

式中，Y_c 为由输入网络得到的三元组，Z 为标准化因子，$Z=Z_1Z_2$，θ 为配置参数，$\theta=(\{\alpha\},\{\mu\})$。

图 4-37 是一个特征分解的例子，在图 4-37 中有五条边，其中两条有已知变量，分别是 $y=1$、$y=0$，另外三条边有未知变量 y。从输入网络可以得到三个三元组，例如 $Y_c=(y_1,y_2,y_3)$ 是其中一个三元组。对于每一条边定义一组特征函数 $f(v_i^s,v_i^u,y_i)$，也可以表示为 $f(x_i,y_i)$。$f(v_i^s,v_i^u,y_i)$ 是一个属性特征函数，可以定义为二值函数，也可以定义为实值函数。针对隐形网络的特点，将其简单定义为一个二值函数，如果在时刻 t 之前用户 v_i^s 转发了用户 v_i^u 的微博，而且用户 v_i^u 反向关注了用户 v_i^s，则可以定义一个特征 $f_j(x_{ij}|=1,y_i=1)$，其值为 1，否则为 0。在三元组特征函数中定义四个特征，两个平衡特征函数和两个不平衡特征函数，如图 4-35 所示。三元组函数也定义为二值函数，如果一个三元组满足结构平衡性质，则其对应的三元组特征函数其值为 1，否则为 0。

模型学习就是要估计参数配置 $\theta=(\{\alpha\},\{\mu\})$，以使对数似然目标函数 $O(\theta)=\log P_\theta(Y|X,G)$ 取得最大值，即：

$$\theta^*=\arg\max O(\theta) \tag{4-27}$$

可以采用牛顿迭代法来解出目标函数，下面以 μ 为例学习得到参数。首先对照目标函数公式（4-27）导出每个 μ_k 的梯度，即：

$$\frac{O(\theta)}{\mu_k}=E[h_k(Y_c)]-E_{p_{\mu_k}(Y_c|X,G)}[h_k(Y_c)] \tag{4-28}$$

式中，$E[h_k(Y_c)]$ 为由数据分布给出的特征函数 $h_k(Y_c)$ 的数学期望，即训练数据中所有三元组特征函数 $h_k(Y_c)$ 的平均值，$E_{p_{\mu_k}(Y_c|X,G)}[h_k(Y_c)]$ 为由估计模型给出的分布 $p_{\mu_k}(Y_c|X,G)$ 的数学期望。

由于三元特征模型的图形结构是任意的，所以直接计算边缘分布难度很大，采用一种称为环路信念传播（Loopy Belief Propagation，LBP）[14]的近似算法来计算边缘分布，LBP 算法具有很好的可实现性和效果。采用 LBP 算法近似估计边缘分布 $p_{\mu_k}(Y_c \mid X,G)$，得到边缘分布的估计值后，通过所有三元组求和得到梯度值。在每次迭代过程中要执行两次 LBP 计算，一次用于估计未知变量 y_i 的边缘分布，另一次用来估计所有三元组的边缘分布。得到梯度值后，按照学习速率 η 来更新每个参数。学习算法如算法 4-2。

算法 4-2　学习算法

输入：网络 G'，学习速率 η

输出：参数 θ 估计值

初始化 $\theta \leftarrow 0$；

重复步骤（1）～（4）直至收敛：

（1）执行 LBP 计算未知变量 $P(y_i \mid x_i,G)$ 的边缘分布；

（2）执行 LBP 计算三元组 c，即 $P(y_c \mid X_c,G)$ 的边缘分布；

（3）根据公式（4-28）计算 μ_k 的梯度值；

（4）按照学习速率 η 更新参数 θ：

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \cdot \frac{O(\theta)}{\theta}$$

得到了参数 θ 的估计值，通过找到一个能够最大化目标函数的配置，即找到 $Y^* = \arg\max O(Y \mid X,G,\theta)$，这样就能够预测未知变量 y_i 的值了。通过再次使用 LBP 算法对含有未知变量 $p_{\mu_k}(y_i \mid x_i,G)$ 的每个关系进行近似估计，计算它们的边缘分布，最后将最大概率值赋给每个关系。

4.5.3　模型验证

下面通过实验数据对社交网络用户关系预测模型进行测试和验证。

1．实验数据集

实验数据来源于新浪微博。该数据集中的每个用户都具有完整的建立连接的历史日志，最终的数据集包括了 94162 个用户，375714 条边，1910748 条微博记录。将每 4 天作为一个时间段，将该数据集分为 15 个时间戳。

2．模型性能对比

实验的目的是预测当用户接收到一个新的关注之后在下一个时间戳内是否会进行反向关注。实验结果显示，60%以上的反向关注行为发生在下一个时间戳内，还有 37%的反向关注行为发生在接下来的三个时间戳内。通过进一步数据分析发现，活跃用户通常在下一

个时间戳内立即进行反向关注或者立即拒绝关注，而一些不太活跃的用户可能不经常登录自己的微博账户，造成反向关注时间的不确定。因此，在实验中使用前 10 个时间戳数据进行训练，使用后 5 个时间戳进行反向关注预测。

与三元特征模型进行对比的方法如下：

（1）支持向量机（SVM）。采用相同的属性作为每条边的特征来训练一个分类模型，然后在测试数据集上进行预测。

（2）逻辑回归分类（LRC）。采用相同的属性作为每条边的特征来训练一个逻辑回归分类模型，然后在测试数据集上进行预测。

（3）条件随机场-平衡（CRF-balance）。采用每条边对应的属性来训练一个条件随机场模型，然后在测试数据集上进行预测。

（4）条件随机场（CRF）。采用所有的属性和结构平衡因子来训练一个条件随机场模型，然后在测试数据集上进行预测。

（5）简化三元特征模型。简化了三元特征模型，不考虑状态同质性和结构平衡性特征。

在上述方法中，SVM、LRC 和 CRF-balance 只考虑了属性因子，而 CRF 虽然采用了结构平衡因子，但是没有考虑无标记数据。而三元特征模型不但包括了所有的因子，同时也考虑到了无标记数据。

为了评价不同方法的预测性能，采用了精确率、召回率、综合评价指标（F_1）和准确率等 4 项指标。其中，精确率是指算法预测中的用户关系数量与数据集中所有用户关系数量之比；准确率也称为查准率，其计算公式为 $P=A/(A+B)$，A 为预测正确的用户关系数量，B 为预测错误的用户关系数量；召回率也称为查全率，其计算公式为 $R=A/(A+C)$，A 为预测正确的用户关系数量，C 为未预测出的用户关系数量；综合评价指标（F_1）是为了平衡准确率和召回，其计算公式为 $F_1=2PR/(P+R)$，F_1 值越高，其综合性能越好。实验结果如表 4-9 所示。

表 4-9　不同方法的反向关注预测性能对比

算法	精确率	召回率	F_1	准确率
SVM	0.7124	0.6007	0.6513	0.9355
LRC	0.7331	0.3081	0.4136	0.9471
CRF-balance	0.9952	0.6024	0.7014	0.9575
CRF	1.0000	0.6678	0.7396	0.9822
简化三元特征	0.9752	0.5806	0.7012	0.9381
三元特征	1.0000	0.8676	0.8995	0.9942

从表 4-9 中可以看到不同方法的预测性能表现，其中三元特征模型的性能明显优于其他四种方法。在综合评价指标上，三元特征模型比 SVM 高出了 27%，比其他三种方法也高出了 20%以上。三元特征模型的优势主要是改进了召回率，它利用了结构平衡相关性

和同质性相关性来检测和处理一些社会学影响。例如，不考虑社会相关性的简化三元特征模型在综合评价指标上性能降低到 70%。另外，三元特征模型还使用了无标记数据，将数据集中的一些潜在相关性也考虑进来。

3．模型性能分析

下面从特征因素影响、算法收敛性、时间跨度影响以及精英用户发现算法影响等 4 个方面对三元特征模型的性能进行分析。

1）特征因素影响分析

在三元特征模型中使用了 5 个特征函数：地理距离（G）、连接同质性（L）、状态同质性（S）、隐形网络相关性（I）和结构平衡相关性（B）。从模型中分别移除每个特征之后得到 5 个简化模型，在这些简化模型上使用综合评价指标来评估其预测性能，测量它们的下降程度，评估结果如图 4-38 所示。

与原有模型相比，简化模型性能下降程度越大，则被移除的特征影响越大，也就是说，该特征对原有模型的预测性能贡献越大。从图 4-38 可以明显看出，与原有模型相比，每个简化模型性能都有所下降，这说明了每个特征对改善预测性能都有贡献，只是贡献大小不同而已。其中移除连接同质性特征后预测性能下降最大，说明该特征对原有模型的预测性能贡献最大。

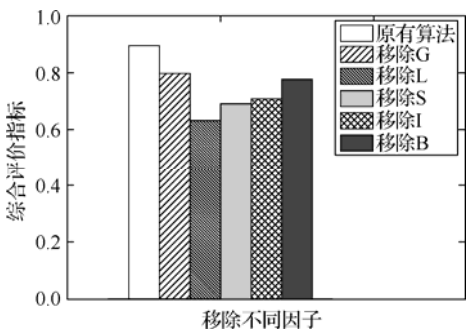


图 4-38　因子贡献分析

2）算法收敛性分析

通过观测实验中 LBP 算法的迭代次数来分析学习算法的收敛性，其实验结果如图 4-39 所示。从图 4-39 可以看出，学习算法能够在迭代 10 次之内收敛，并且在迭代 7 次之后，三元特征模型的预测性能变得稳定，这表明学习算法非常有效，而且具有很好的收敛性。

3）时间跨度影响分析

图 4-40 显示了在不同时间跨度设置上预测性能的综合评价指标值，当时间跨度设置为 1 个时间戳或 2 个时间戳时，模型的预测性能急剧下降，当时间跨度设置为 3 个时间戳

时，模型的预测性能达到 90%。图 4-40 中超过 90%的反向关注行为发生在前 3 个时间戳内，而发生在前两个时间戳内的反向关注行为不到 80%。

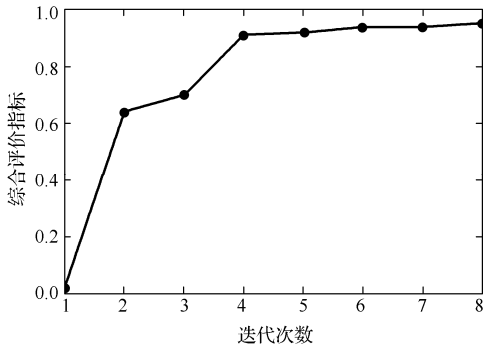


图 4-39　学习算法的收敛性分析

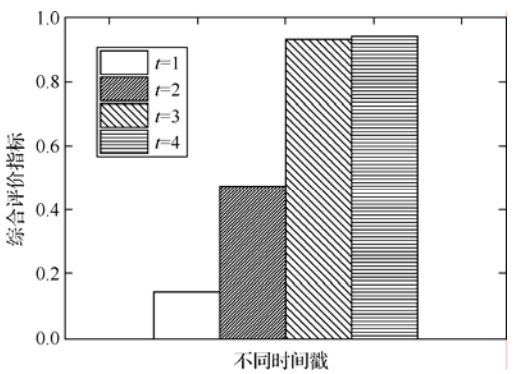


图 4-40　不同时间跨度的反向关注预测

4）精英用户发现算法影响分析

状态同质性特征来自于精英用户发现的结果。在发现精英用户时分别使用了三种不同的算法：PageRank、度计算和 (α,β) 算法。表 4-10 显示了不同精英用户发现算法的三元特征模型反向关注预测性能评估结果。从表 4-10 中可以得到与图 4-33 相同的结果，即使用 (α,β) 算法来发现精英用户，其模型的反向关注预测性能最好，这说明了模型中的状态同质性特征是有效的。

表 4-10　不同精英用户发现算法的反向关注预测性能对比

发现算法	精确率	召回率	综合评价	准确率
(α,β)	1.0000	0.8746	0.9351	0.9932
PageRank	1.0000	0.8017	0.9013	0.9725
度计算	1.0000	0.7674	0.8528	0.9862

以上的实验结果表明，三元特征模型融入社会学理论，能够明显地提高模型的双向关系预测性能。

4.6　社交网络意见领袖识别

意见领袖又称舆论领袖，是指在信息传播网络中经常发表意见并具有相当影响力的“活跃分子”，他们在信息制造和传播过程中发挥着重要的作用，由他们将信息传播给受众，在意见领袖的引导和影响下，局部意见可能演化为网络舆论。

统计数据显示，很多网民并不直接发表意见，而是通过关注和转发意见领袖的信息来表达自己的态度和倾向性，即所谓服从权威现象。通过意见领袖发表引导性信息来影响网民，能够有效地触发整个网络舆论的影响力。因此，意见领袖在推动信息传播、引导网络舆论中发挥着重要的作用。

随着网络舆论影响力的不断加大，人们提出了不同网络信息交互平台（如网络论坛、社交网络等）的意见领袖识别方法。下面主要研究社交网络意见领袖识别方法。

4.6.1　识别方法

在社交网络中，种子节点的影响力对推动信息传播是非常重要的。一些通过病毒式市场营销方式来推销其产品、服务的公司或用户对如何选择具有影响力的种子节点怀有很大的兴趣。比如 A 公司想在社交网站为其产品做广告，由于广告费用有限，只能投放 K 个用户，A 公司希望这些最初的用户能够喜欢其产品，并以他们作为种子节点，在社交网络中以口碑相传方式来影响他们的朋友，让他们的朋友也喜欢其产品，而他们的朋友又通过社交网络进一步影响更多的朋友，使更多的用户都能喜欢其产品。A 公司当然希望最初选择的用户（即种子节点）都具有较大影响力，所影响的人数尽可能地多，从而以最少的费用达到最大的广告效益。可见，种子节点在信息传播过程中发挥了重要的作用，他们相当于意见领袖，通过他们的引导和影响，能够有效地推动信息传播。

因此，可以通过搜索影响力最大的种子节点来识别意见领袖，并将此类问题归结为影响力最大化问题。

在描述算法之前，首先定义影响力最大种子节点搜索问题。

定义 $\sigma(\cdot)$ 为影响力函数，S 为种子节点集合，U 为搜索节点集合。$\sigma(S)$ 表示种子节点集合 S 的影响力。

如果对于任何元素 x，$y \in RK$ 有 $f(x \vee y)+f(x \wedge y) \leqslant f(x)+f(y)$，则函数 $f: R^k \to R$ 是子模函数。

由此可以得出如下结论：

（1）如果 f 是子模函数，则 $\forall A \subset B \subset N$，$\forall j \in N \setminus B$，则有 $f(A+j)-f(A) \geqslant f(B+j)-f(B)$，任何子模函数都具有单调、非负等性质。

（2）在独立级联模型、带权级联模型和线性阈值模型的任何一个实例中，影响力函数 $\sigma(\cdot)$ 是一个子模函数。

随着集合 S 的节点数目增多，集合 U 中所有节点的影响力都在逐渐减弱，具有单调递减性。

根据节点影响力所具有的子模函数性质，可以采用贪婪算法来搜索种子节点。为了克服经典贪婪算法计算效率较低的问题，文献[15]提出一种优化贪婪算法（CELF）来搜索种子节点，其搜索过程分为两步：首先计算所有节点影响力，选择影响力最大的节点作为第一个种子节点加入到种子节点集合中。然后选择余下的种子节点，在每次选择种子节点过程中，根据影响力具有子模函数性质，算法只计算部分影响力大的节点。

与经典贪婪算法相比，CELF 算法的计算效率有了较大的提高，但是在大规模社交网络中搜索种子节点仍然非常耗时，需要对搜索算法做进一步的改进。

大量的研究表明，社交网络的节点度呈幂律分布，并且节点度与其影响力存在强关联性，这意味着社交网络中存在着大量影响力较小的节点和少量影响力较大的节点，而种子节点则是影响力较大的节点。对于大量影响力较小的节点，即节点度数较小的节点，成为种子节点的概率非常低，可以考虑将它们排除在种子节点搜索范围之外，从而缩小种子节点搜索的范围。

基于以上事实，对搜索算法做进一步的改进，改进后的搜索算法称为高节点度贪婪算法（HD_Greedy），其基本思想是在极小部分高度数节点中搜索种子节点，其搜索过程分为如下步骤：

（1）输入高度数节点占所有节点的百分比 r，$0<r\leqslant 1$，对所有节点按度数由大到小排序，选择排序前 r 的节点构成一个高度数节点集合。

（2）在高度数节点集合中搜索种子节点，计算其影响力，按影响力由大到小排序，加入到种子节点集合中，直到所有节点搜索完毕。

这样，在种子节点集合中，形成一个按影响力大小排序的种子节点序列，选择影响力最大的节点作为第一个种子节点。

4.6.2 算法验证

下面通过实验数据对社交网络意见领袖识别算法性能进行测试和验证。

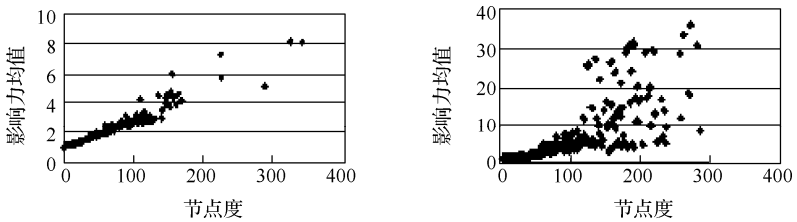
1．实验数据集

实验数据来源于论文共享网站 arxiv（www.arxiv.org），其中网络中的节点代表学者，

边代表学者间科研合作关系，而科研合作关系主要体现在论文合作方面。第一个科研合作网来自于“高能物理理论”版块，从 1991 年至 2003 年，用 NetHEPT 表示，它包含 15233 个节点和 58891 条边。第二个科研合作网来自于“物理学”版块，用 NetPHY 表示，它包含 37154 个节点和 231584 条边。

2．节点度与影响力关联性

对于节点度与影响力的关联性，采用独立级联信息传播模型进行分析。图 4-41 给出了节点的度与影响力散点图，其中横坐标表示节点度大小，纵坐标表示节点的影响力均值。



（a）NetHEPT中节点度与影响力散点图　　（b）NetPHY中节点度与影响力散点图

图 4-41　独立级联模型中节点的度与影响力散点图

从图 4-41 可以看出，节点度数越大，影响力也就越大，但是在度数较大节点中影响力均值偏差较大，原因是由于度数较大节点数量少，个别节点的影响力对均值影响较大。例如，在 NetPHY 中所有度数大于 150 的节点只有 157 个，其平均在每个节点度的分布约为 2.5 个，所以节点的度数与影响力存在着较强的关联性。这也符合实际的社交网络。例如，在人际关系网中，用户的朋友关系越多，受他影响的人数也就越多，其深层次的原因是社交网络中的边是信息传播的唯一路径，高度数节点连接着大量的边，因而其影响的范围相对也就较大。

3．算法性能对比

下面是 HD_Greedy 算法与其他经典算法的性能对比实验，实验方法是各个算法分别在 NetHEPT 和 NetPHY 中计算种子节点的影响力及其耗费的时间。

由于当 r = 1%时 HD_Greedy 算法性能最优，因此实验将取 r = 0.01 的高节点度贪婪算法（HD_Greedy_01）分别在独立级联模型、带权级联模型和线性阈值模型等三个不同信息传播模型上与其他算法进行对比实验，算法包括 Greedy 算法（即 CELF 算法）、Degree 算法、NewGreedy 算法和 MixGreedy 算法等。独立级联模型的节点影响因子 p 取值为 0.01。

种子节点的数目为 50。为了确保节点影响力计算的精确性，每个算法对节点的影响力计算 20000 次，取平均值作为节点最终影响力，以防止随机概率引起的误差。由于

Degree 算法在寻找种子节点过程中只选取度数最大的节点，不需计算节点的影响力，它的计算时间非常短，约为 0.004 秒，比其他算法快 6 个数量级以上，所以在各个算法的计算时间对比时，Degree 算法不参与比较。

1）独立级联模型对比

图 4-42 为各个算法在独立级联模型中的比较，其中纵坐标为影响力平均值，横坐标为种子节点数目。从图 4-42 可以看出，在 NetHEPT 和 NetPHY 中，HD_Greedy_01 算法与 Greedy、NewGreedy 和 MixGreedy 算法所得到的种子节点影响力增长曲线几乎重叠，而且最终的影响力相差不到 1%，明显高于 Degree 算法，尤其在 NetPHY 中，相差了 17%。这说明 HD_Greedy_01 算法得到的种子节点影响力与 Greedy、NewGreedy 和 MixGreedy 算法相近，而明显高于 Degree 算法。从图 4-42 还可以看出，HD_Greedy_01 算法得到的种子节点影响力增长曲线与 Greedy 算法几乎完全重叠，最终影响力相差很小，这说明所有种子节点几乎都集中在部分度数较高的节点集合中，HD_Greedy_01 算法只在少部分高度数节点集合中搜索种子节点并不会损失种子节点影响力。

图 4-43 为独立级联模型下的算法计算时间对比情况，时间单位为秒（s）。从图 4-43 可以看出，在 NetHEPT 和 NetPHY 中，HD_Greedy_01 算法与 Greedy 算法计算相比，时间分别缩短了 75%和 64%，效率大为提高，而且也明显低于 MixGreedy 和 NewGreedy 算法的计算时间。实验发现，在 NetHEPT 和 NetPHY 中，NewGreedy 和 MixGreedy 算法计算时间不稳定，NewGreedy 算法在 NetPHY 中的计算时间比 Greedy 算法快了近 60%，而在 NetHEPT 中的计算时间甚至比 Greedy 算法还要慢，同样的情况也出现在 MixGreedy 算法中，计算时间的不稳定性不利于算法的性能评估。HD_Greedy_01 算法不仅大大缩短了计算时间，而且具有良好的稳定性。因此，在独立级联模型的算法对比实验中，HD_Greedy_01 算法性能更好。

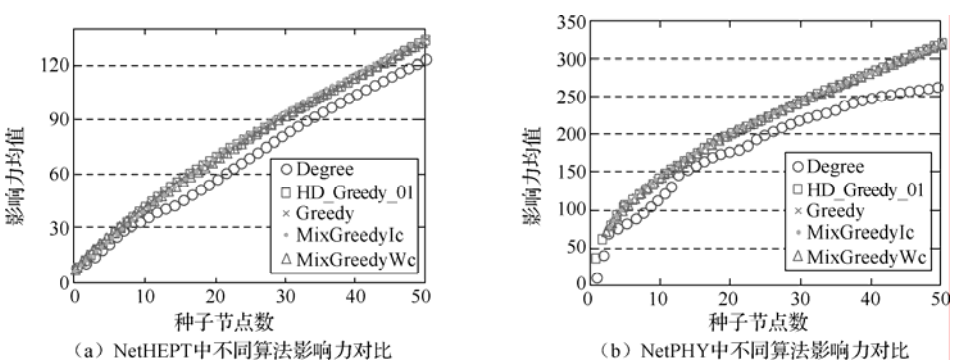


图 4-42　独立级联模型中不同算法影响力对比

2）带权级联模型对比

图 4-44 给出了各个算法在带权级联模型中的比较，其中纵坐标为影响力平均值，横坐标为种子节点数目。从图 4-44 可以看出，相比于独立级联模型，在带权级联模型中种子节点影响力增大，但 HD_Greedy_01 算法与 Greedy、NewGreedy 和 MixGreedy 算法得到的种子节点影响力的增长曲线也几乎是重叠的，各个贪婪算法得到的种子节点影响力相差也不大，不到 1.5%。而 Degree 算法得到的种子节点影响力明显不及贪婪算法，在 NetHEPT 和 NetPHY 中分别仅有贪婪算法的 85.5%、62.3%。另外，随着网络节点数目增多，Degree 算法得到的种子节点影响力与贪婪算法的差距也在增大，这说明 Degree 算法不适合在大规模社交网络中求解种子节点影响力。

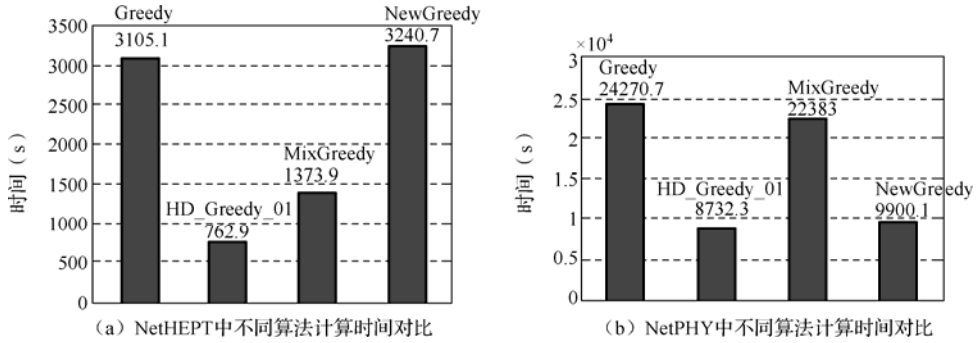
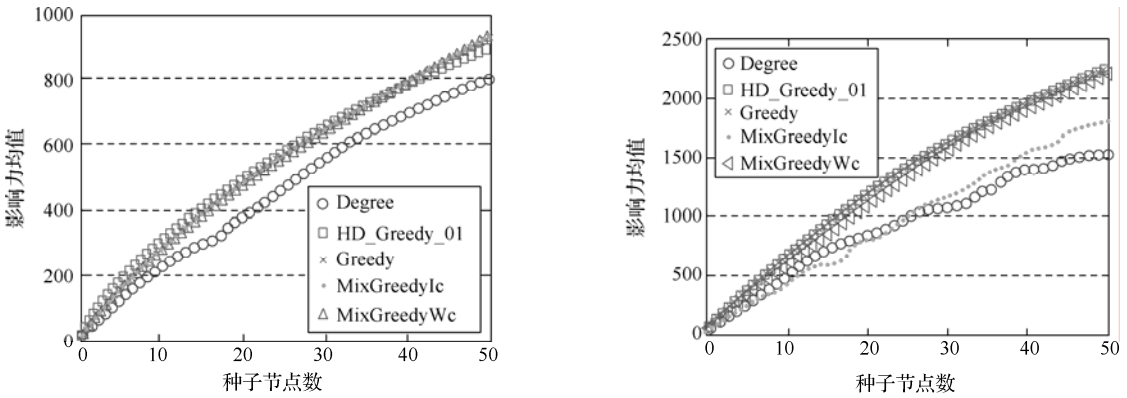


图 4-43　独立级联模型中不同算法计算时间对比

图 4-45 是在带权级联模型下各个算法计算时间对比情况，时间单位为秒（s）。从图 4-45 可以看出，在 NetHEPT 中，HD_Greedy_01 算法比 Greedy 算法的计算时间缩短了 48.5%，而在 NetPHY 中计算时间缩短了 55.6%。同时它还是速度最快的算法，比次快的 MixGreedy 算法在 NetHEPT 和 NetPHY 中分别快了 44.5%和 11.2%。NewGreedy 算法的计算时间比较长，比 Greedy 算法还长，在 NetHEPT 中的计算时间是 Greedy 算法的 244%，在 NetPHY 中则达到了 275%。

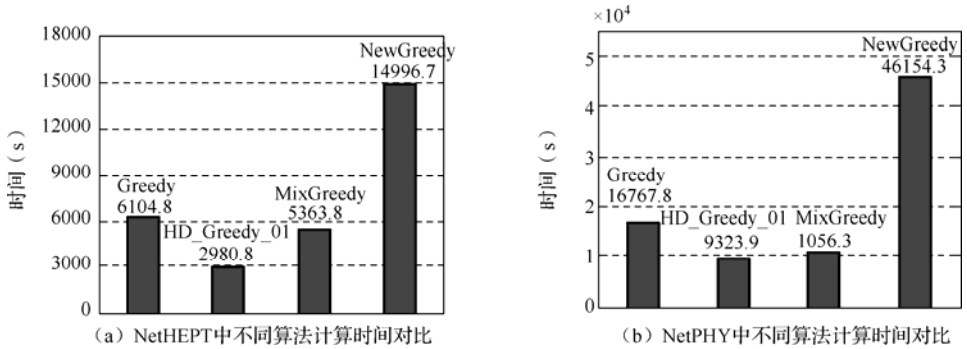
3）线性阈值模型对比

由于 NewGreedy 和 MixGreedy 算法是基于独立级联模型提出的，仅适应于独立级联模型和带权级联模型。为了扩展到线性阈值模型中，首先在独立级联模型和带权级联模型中用 MixGreedy 算法求得种子节点，然后在线性阈值模型中计算种子节点的影响力。图 4-46 给出了线性阈值模型中不同算法的影响力对比，其中纵坐标为影响力平均值，横坐标为种子节点数目，MixGreedyIc 和 MixGreedyWc 分别表示在独立级联模型和带权级联模型下 MixGreedy 算法得到的结果。



（a）NetHEPT中不同算法影响力对比　　（b） NetPHY中不同算法影响力对比

图 4-44　带权级联模型中不同算法的影响力对比



（a）NetHEPT中不同算法计算时间对比　　（b）NetPHY中不同算法计算时间对比

图 4-45　带权级联模型中不同算法计算时间对比

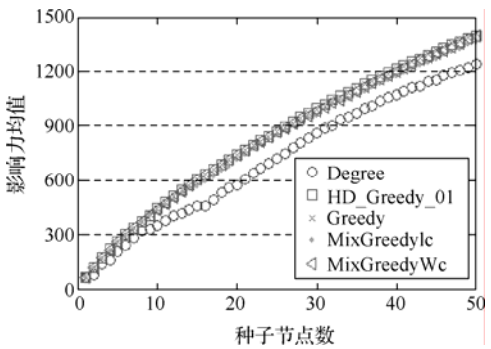


图 4-46　线性阈值模型中不同算法的影响力对比

从图 4-46 可以看出，在线性阈值模型中，HD_Greedy_01 算法与 Greedy 算法的种子节点影响力增长曲线几乎是重叠的，MixGreedyWc 算法得到的种子节点影响力与它们相

差也很小。Degree 算法在 NetPHY 中得到的种子节点影响力仅为 Greedy 算法的 67.3%。MixGreedyIc 算法在 NetPHY 中得到的种子节点影响力与 HD_Greedy_01、Greedy 和 MixGreedyWc 三个算法相差比较大，达到了 23.9%，而在独立级联模型中 MixGreedy 算法得到的种子节点影响力与其他贪婪算法相差很小。这也从侧面反映了 MixGreedy 算法在独立级联模型或带权级联模型中得到的种子节点并不能在线性阈值模型中得到很好的扩展。

以上的实验结果可以看出，在不同的信息传播模型中，HD_Greedy 算法得到的种子节点影响力与其他贪婪算法接近，但计算效率有了较大提高。尤其在大规模社交网络中，它的计算效率更高，这表明 HD_Greedy 算法更适应于大规模社交网络。虽然 NewGreedy 和 MixGreedy 算法得到的种子节点影响力与 CELF 算法非常接近，但存在着计算时间不稳定以及在线性阈值模型中可扩展性较差等问题。而 Degree 算法得到的种子节点影响力与所有贪婪算法相差较大。因此，HD_Greedy 算法性能更优。

另外，在 HD_Greedy 算法中，高度数节点占所有节点百分比 r 值的设置对算法性能有较大的影响，r 值设置过小，算法得到的种子节点影响力将受到损失；r 值设置过大，算法的计算效率将大大降低；当 $r = 1$ 时，它完全蜕变为 CELF 算法。r 值的确定与许多参数是相关联的，如社交网络中节点数量及拓扑结构、信息传播模型的影响因子大小以及种子节点数量等。

参考文献

[1] D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network[C]. In Proc. of 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.

[2] Bass Frank. A new product growth model for consumer durables[J]. Management Science, 1969, 15 （5）: p215-227.

[3] Zanette D H. Dynamics of rumor propagation on small-world networks[J]. Physical Review E, 2002, 65: 041908.

[4] Moreno Y, Nekovee M, Pacheco A F. Dynamics of rumor spreading in complex networks[J]. Physical Review E, 2004, 69: 066130.

[5] M. Granovetter. The strength of weak ties[J]. The American Journal of Sociology, vol. 78, no. 6, pp. 1360-1380, 1973.

[6] S. Milgram. Behavioral study of obedience[J]. Journal of Abnormal and Social Psychology, vol. 67, no. 4, pp. 371-378, 1963.

[7] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks[C]. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM,

2008: 7-15.

[8] Onnela J P, Saramäki J, Hyvönen J, et al. Structure and tie strengths in mobile communication networks[J]. Proceedings of the National Academy of Sciences, 2007, 104（18）: 7332-7336.

[9] May R M, Lloyd A L. Infection dynamics on scale-free networks[J]. Physical Review E, 2001, 64（6）: 066112.

[10] Cha M, Mislove A, Gummadi K P. A measurement-driven analysis of information propagation in the flickr social network[C]. Proceedings of the 18th international conference on World wide web. ACM, 2009: 721-730.

[11] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the web[J]. 1999.

[12] He J, Hopcroft J, Liang H, et al. Detecting the structure of social networks using （α, β）-communities[M]. Algorithms and Models for the Web Graph. Springer Berlin Heidelberg, 2011: 26-37.

[13] Sandhu R S, Coyne E J, Feinstein H L, et al. Role-based access control models[J]. Computer, 1996, 29（2）: 38-47.

[14] Murphy K P, Weiss Y, Jordan M I. Loopy belief propagation for approximate inference: An empirical study[C]. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999: 467-475.

[15] J Leskovec, A Krause, C Guestrin, C Faloutsos, J Van Briesen. Cost-effective outbreak detection in networks[C]. In Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 420-429, 2007.

第5章

微博网络用户转发模型

5.1 引言

在 1.4 节中，对微博网络及其信息传播模式进行了介绍。从中可以看出，微博网络作为一种特殊的社交网络，用户不但可以有选择地连接感兴趣的用户，关注其信息，而且也可以被其他用户相互连接，交流信息，具有社交网络和媒体网络的双重特性。微博网络给大众提供了一个自由发表意见并与他人分享的平台，推动了公众话语权的回归，开创了一个平民化的信息传播模式。同时，也带来了与社交网络相类似的信息安全方面的挑战，包括进行非法联络、传播谣言、煽动闹事等，容易引起社会群体事件。

由于微博网络在当今社会信息传播中发挥越来越重要的作用，因此对微博网络信息传播机制的研究具有重要的现实意义，通过分析用户行为及爱好，预测用户行为演化趋势，不仅可以为网络舆情监控、突发事件预测等提供科学依据，还可以为商家分析用户购买喜好、产品推荐以及精准投放广告等提供帮助。

微博转发是微博网络提供的一种信息传播机制，用户可以将关注者发布的微博转发到自身平台，然后分享给粉丝。通过这种信息传播机制，使得微博能够在更大范围内传播和共享。因此，用户转发行为是推动微博网络信息传播的重要因素。

本章主要对微博网络的用户转发行为及预测、微博转发特性及预测、微博转发峰值等问题进行分析和研究，有助于认识微博网络信息交流和互动的内在动力和规律。

5.2 微博用户转发特性

用户转发是微博网络最主要的信息传播特性，用户转发行为与用户关系类型有关。在微博网络中，用户之间存在三种社会纽带关系，即强连接、弱连接及权威连接，不同的社会纽带关系对用户转发行为的影响也不同。因此，在分析用户转发行为时，首先需要识别用户之间的社会纽带关系。

在识别社会纽带关系时，首先需要提取出权威比率、微网络结构、地理距离以及性别等特征，然后采用适当的模型来分析各个特征间的相关性，并根据相关性来分析用户转发行为的内在动力。

5.2.1 微博用户转发行为特性

在微博网络中，对于信息的传播，每个用户扮演两种不同的角色，一个是接收者，转发来自于其他用户发布的微博；另一个是发布者，用户发布自己的微博。

对于微博用户转发行为，主要从接收者角度来分析，也就是说，主要考虑用户是否会转发其他用户的微博，而不考虑自己的微博是否被其他用户所转发。例如，对于用户 A 来说，仅考虑用户 A 是否会转发其关注者 B 的微博，而不考虑用户 A 发布的微博是否会被其他用户或粉丝转发。从本质上来说，从接收者或发布者的角度来研究微博转发行为是一样的，只是用户角度不同而已，从接收者角度来研究只是为了简化问题。

在微博网络中，定义 $A \rightarrow B$ 为一条关注边，用户 A 关注用户 B（或者用户 B 被用户 A 所关注）。在该关注边中，用户 A 是用户 B 的粉丝或用户 B 是用户 A 的关注者。成为用户的粉丝意味着能够自动地接收该用户发布的微博。因此，在关注边 $A \rightarrow B$ 中，用户 A 能够接收到用户 B 的所有微博。同时用户 A 可以转发用户 B 的微博，通过这一功能，用户 A 的粉丝能够自动地接收到该微博，并可以再次转发。因此，转发功能实现了微博信息在网络中的传播。但是网络关注边方向与微博传播的方向是相反的，例如在 $A \rightarrow B$ 中，表示用户 A 关注用户 B，而微博则是由用户 B 转发到用户 A。

图 5-1 给出微博用户关注网络图，Bob 关注了 Greg、Harry、Fred 以及 Carol 等，同时被 Dave、Alice、Greg 以及 Harry 等关注，如图 5-1 中的不同箭头方向，其中存在一部分用户与 Bob 相互关注，例如 Greg 与 Harry。对于 Bob 而言，是否会转发被他关注的用户微博。

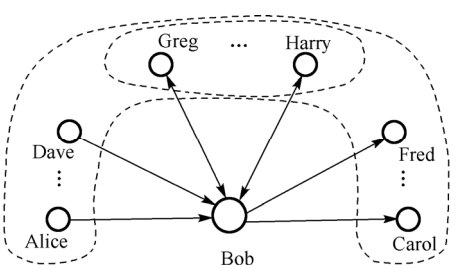


图 5-1 微博网络用户示例图

这里的一个核心问题是对于关注边 $A \rightarrow B$，什么因素导致用户 A 转发用户 B 的微博。由于用户之间存在着不同类型的社会纽带关系，不同类型的社会纽带关系导致不同的微博转发行为，因此需要从社会纽带关系来研究微博转发行为。

在社交网络中，用户之间存在三类社会纽带关系：强连接、弱连接和权威连接。在第 4 章中详细论述了社交网络的强连接与弱连接。在微博网络中，用户关系除了存在着强连接与弱连接外，还存在另一种社会关系，即权威连接关系，例如一个普通的用户连接一个名人，它完全不同于强连接和弱连接，主要表现在用户之间的非对称性，非对称性包括两个方面，一是用户影响力的非对称，例如名人的影响力比一般用户大很多；二是信息传播的非对称性，例如名人的微博很容易被一般用户转发，而一般用户的微博很难被权威用户转发。在权威连接关系中，信息传播方向一般由权威高的用户到权威低的用户。在社会科学中，这种现象称为服从权威[1]。

显然，用户之间不同的社会纽带关系将导致不同的微博转发行为。但是在微博网络中，用户之间的社会纽带关系是很难发现的，也很难定义用户之间是否存在强连接、弱连接或权威连接等关系。因此，如何识别出不同的社会纽带关系是一个关键性问题。

文献[2～5]研究表明在社交网络中信息传播存在两个重要进程，即同化与社会影响。同化是指信息在网络传播过程中容易导致用户与自己观点、价值观相似的用户建立连接关系，最终使社交网络的结构发生变化。社会影响则是指信息在网络传播过程中导致相邻用户的观点、价值观等属性逐渐趋于一致，最终使两个用户具有相似性。因此，在信息传播过程中，不同社会纽带关系的用户将受到同化和社会影响的影响，最终导致了网络结构和用户属性的变化，也就是说，不同的社会纽带关系将具有不同的网络结构和用户属性。反过来说，通过提取网络结构和用户属性等特征，就能够识别出用户之间的社会纽带关系。网络结构特征可以从微博用户关注图得到，包括权威比率、微网络结构；而用户属性则可以从用户的个人资料中提取出来。

1．权威连接比率

在社交网络中，关注边 $A \to B$ 上的两个用户在不同的社会纽带关系中有着不同的社会权威。在强连接或弱连接中，用户大多是朋友、同事等对等关系，社会地位是平等的。而在权威连接中，名人与普通用户的社会地位并不平等，名人往往比普通用户有更高的社会地位和权威。因此，关注边中不同的社会权威比率可以间接地反映出两个用户是否存在着强连接、弱连接或权威连接等关系。由于一个用户的粉丝数反映了用户的影响力，因此可以采用用户的粉丝数来刻画用户权威。对于关注边 $A \to B$，两个用户权威比率定义如下：

$$P_{A \to B} = \frac{\#\text{Follower}(B)}{\#\text{Follower}(A)} \tag{5-1}$$

公式（5-1）反映了关注边 $A \to B$ 中两个用户的权威差异性，$P_{A \to B}$ 值越大，表明两个用户的社会地位越不平等，则关注边 $A \to B$ 为权威连接的可能性越大。

$P_{A\to B}$是一个连续值，需要对其离散化，信息熵是最常用的离散化度量方法。基于熵的离散化是一种监督的、自顶向下的分裂方法，在计算和确定分裂点（划分属性区间的数据值）时利用类分布信息，选择具有最少熵的属性作为分裂点，使区间划分离散化。

为了解释基于熵的离散化的基本思想，必须考查一下分类。假定需要根据属性 A 和某个分裂点上的划分将 D 中的元组分类。理想地，希望该划分导致元组的准确分类。例如，如果有两个类，希望类 C_1 的所有元组落入一个划分，而类 C_2 的所有元组落入另一个划分。然而，这是不可能的。因为第一个划分可能包含许多 C_1 的元组，但也包含某些 C_2 的元组，而在第二个划分中情况则相反。在该划分之后，为了得到完全的分类，还需要多少信息？假定 $D_1,D_2,\cdots,D_m$ 为 $P_{A\to B}$ 连续值依次划分成的 m 个元组子集，则该划分对 D 的元组分类的期望信息需求，由公式（5-2）给出。

$$\text{Infor}(D)=\sum_{i=1}^{m}\frac{|D_i|}{|D|}\text{Entropy}(D_i) \tag{5-2}$$

式中，$|D_i|$为第 i 个子集中元组的个数，$|D|$为 D 中元组的个数。且给定子集 i 的熵函数 $\text{Entropy}(D_i)$ 根据集合中元组的类分布来计算。微博转发元组包括两个类：转发与不转发，分别用 C_1，C_2 表示，则 D_i 的熵是：

$$\text{Entropy}(D_i)=-p_1\log_2(p_1)-p_2\log_2(p_2) \tag{5-3}$$

式中，p_1 为 D_i 中转发元组类的概率，由 D_i 中 C_1 类的元组数除以 D_i 中的元组总数$|D_i|$来确定；p_2 为 D_i 中不转发元组类的概率，概率计算类似于 p_1。

在选择分裂点时，希望选择产生最小期望信息需求［即 $\min(\text{Infor}(D))$］，这将导致在划分 m 个子集之后，对元组完全分类所需的期望信息量最小，等价于具有最大信息增益的属性。

将 $P_{A\to B}$ 离散化为三大类：Small、Medium 和 Large。对 $P_{A\to B}$ 信息熵离散化后结果如表 5-1 所示。

表 5-1　权威比率离散化分布

属性	值
Small	$P_{A\to B}<4$
Medium	$4<P_{A\to B}<100$
Large	$100<P_{A\to B}$

2．微网络结构

在微博网络中，用户关注关系将构成一个有向网络图，而具有不同社会纽带关系的用户有着不同的网络结构，这里主要关注用户间微网络结构。相比于全局网络结构，微网络

结构更能够反映出用户之间的社会纽带关系。例如，在权威连接中，用户与其关注者往往是单向关系，通常是普通用户单向地关注权威人士而权威人士并不会关注普通用户。然而在强连接或弱连接中，用户与其关注者更容易相互关注，形成双向关系。为了区分双向关注边 $A \leftrightarrow B$ 是否为强连接或弱连接，引入第三方邻居用户，即是否存在与用户 A、B 都相互关注的用户。如果用户 A、B 间的社会纽带关系越强烈，则与他们都相互关注的用户数就越多。例如，在朋友圈子中，两个用户关系越紧密，则他们共同拥有的朋友数量就越多。

下面通过三种不同类型的微网络结构来区分不同类型的社会纽带关系，如图 5-2 所示。在 Pattern I 中，用户 A、B 相互关注，同时至少存在一个第三方用户与用户 A、B 也两两相互关注，即存在强连接的三角关系；在 Pattern II 中，用户 A、B 相互关注，与 Pattern I 不同之处在于，不存在一个第三方用户与用户 A、B 两两相互关注；在 Pattern III 中，用户 A、B 是单向关注关系。这三种不同类型的微网络结构反映了不同的社会纽带关系，强连接的用户容易形成 Pattern I 微网络结构，弱连接的用户容易形成 Pattern II 微网络结构，而权威连接的用户则容易形成 Pattern III 微网络结构。

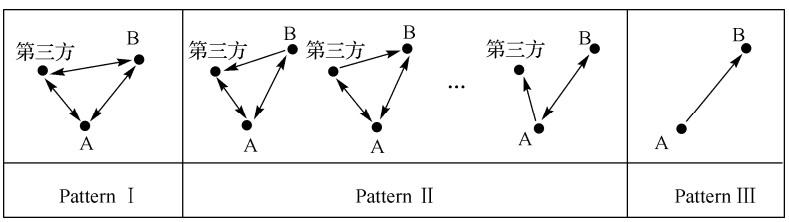


图 5-2　三种不同类型的微网络结构

3．地理距离

用户之间居住的地理距离对用户的社会纽带关系也有着不同的影响，因此地理距离也可以用来作为区分社会纽带关系的特征。例如，在强连接关系中，经常见面相处、距离较近的两个用户，容易建立起这种关系。如果两个用户距离较远，维持这种关系的成本将会大大增加，而不容易建立起这种关系。因此，建立强连接关系常常受到地理距离的限制，距离较近的两个用户比距离较远的两个用户更加容易建立强连接关系。而在权威连接关系中，普通用户可以通过一些公共媒体平台单向地关注权威人士，维持这种关系较为容易，由此可见，地理距离对权威连接的建立影响较小。

利用用户资料中所填写的省份与城市来定义两个用户的地理距离的远近，细分为三类：Near、Medium 和 Far，它们分别被定义为关注边 $A \rightarrow B$ 中两个用户处在相同城市、处在不同城市但处在相同的省份、处在不同的省份。

4．用户性别

用户之间不同的性别也有可能预示着不同的社会纽带关系。例如，相对于异性，同性关系的用户之间更加容易建立朋友关系。另外，社会科学家研究也表明，不同性别用户的思想观念、价值取向以及思维方式等并不相同，因此对于相同的微博，他们的兴趣也不一样，需要考虑不同性别对转发行为的影响。对于关注边的两个用户，可以分为四类：MM、MF、FM 和 FF，其中 MM 表示男性用户 A 关注男性用户 B；MF 表示男性用户 A 关注女性用户 B；FM 表示女性用户 A 关注男性用户 B；FF 表示女性用户 A 关注女性用户 B。

以上所有特征都将影响着用户的微博转发行为，表 5-2 给出所有特征及相应的值，其中圆括号中的索引值代表相应特征值，例如，在权威比率特征中，索引值 1、2、3 分别代表相应的 Small、Medium 和 Large 值，其他类似。

表 5-2　微博转发特征以及相应值

	属性	值
A	权威比率	Small，Medium，Large （1，2，3）
B	微网络结构	Pattern I， Pattern II， Pattern III（1，2，3）
C	地理距离	Near， Medium，Far （1，2，3）
D	性别	MM，MF，FM，FF （1，2，3，4）
E	转发	Yes， No（1，0）

5.2.2　转发行为特性分析模型

Log-linear 模型[6]是一个分类变量及其相互关系的统计模型，通过类似于传统的回归方法来评估和建模。下面应用 Log-linear 模型来分析各个特征与微博转发因子权重。

在 N 维分类变量中，对于变量 $y_{i_1 i_2 \cdots i_N}$，i_r 是第 r 个分类变量的值，定义期望值 $\hat{y}_{i_1 i_2 \cdots i_n}$ 的对数是线性函数，即：

$$\hat{l}_{i_1 i_2 \cdots i_N} = \log_{\hat{y}_{i_1 i_2 \cdots i_N}} = \sum \gamma^G (i_r \mid d_r \in G) \tag{5-4}$$

式中，γ 为 Log-linear 模型中变量系数，系数对应于任何一个子集，而子集都是从更高层次的集合抽取出来的。

为了清楚地描述 Log-linear 模型，给定一个 5 维的分类变量，维度分别为 A（1，…，L）、B（1，…，J）、C（1，…，K）、D（1，…，L）和 E（1，…，M），则饱和的 Log-linear 模型如下：

$$\begin{aligned} \log_{\hat{y}_{ijklm}} = {} & \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D + \gamma_m^E + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} \\ & + \gamma_{il}^{AD} + \gamma_{im}^{AE} + \gamma_{jk}^{BC} + \gamma_{jl}^{BD} + \gamma_{jm}^{BE} + \gamma_{kl}^{CD} \end{aligned} \tag{5-5}$$

$$
\begin{aligned}
&+\gamma_{km}^{CE}+\gamma_{lm}^{DE}+\gamma_{ijk}^{ABC}+\gamma_{ijl}^{ABD}+\gamma_{ijm}^{ABE}+\gamma_{ikl}^{ACD}\\
&+\gamma_{ikm}^{ACE}+\gamma_{ilm}^{ADE}+\gamma_{jkl}^{BCD}+\gamma_{jkm}^{BCE}+\gamma_{jlm}^{BDE}\\
&+\gamma_{klm}^{CDE}+\gamma_{ijkl}^{ABCD}+\gamma_{ijkm}^{ABCE}+\gamma_{ijlm}^{ABDE}\\
&+\gamma_{iklm}^{ACDE}+\gamma_{jklm}^{BCDE}+\gamma_{ijklm}^{ABCDE}
\end{aligned}
$$

饱和模型包括了一个 5 阶因子共效系数，所有的 4 阶、3 阶、2 阶因子共效系数，所有的单个因子系数以及一个均值系数 γ。其中，单个因子 γ_i^A 表示分类变量 A（1，…，I）的第 i 个值的权重大小，2 阶因子 γ_{ij}^{AB} 表示分类变量 A（1，…，L）第 i 个值与分类变量 B（1，…，J）的第 j 个值共效后的因子权重，其他系数相类似。

所有因子系数都有一定的限制，即分量变量系数之和为零，即：

$$
\begin{aligned}
&\sum_i \gamma_i^A=\sum_j \gamma_j^B=\cdots=\sum_m \gamma_m^E=0\\
&\sum_i \gamma_{ij}^{AB}=\sum_j \gamma_{ij}^{AB}=\cdots=\sum_m \gamma_{lm}^{DE}=0\\
&\cdots\cdots\\
&\sum_i \gamma_{ijklm}^{ABCDE}=\sum_j \gamma_{ijklm}^{ABCDE}=\cdots=\sum_m \gamma_{ijklm}^{ABCDE}=0
\end{aligned}
\qquad (5\text{-}6)
$$

在该模型中，γ 系数用来表示各个分类变量相互作用的权重大小，γ 系数越大，则分类变量间相互作用越显著。因此，γ 系数值可以用来区分分类变量间相互作用的强弱。

由于饱和 Log-linear 模型包括了所有的系数，其中包括一些很弱的系数，而这些很弱的系数对模型作用较小，可以忽略不计，若保留它们有可能导致数据过分拟合。因此需要去除这些较弱的系数，选择一个精简的模型来拟合数据。精简的模型需要平衡两个方面的目标，一是该模型应当足够的复杂，能够很好拟合数据；二是该模型又要尽量的简练，能够很好地解析数据，防止数据过分拟合现象。这两个目标是一个相互平衡的过程，一个理想的模型需要同时考虑以上两个方面的目标。

可以选择不同的模型来拟合数据，例如在 k 个分类变量中，则存在 2^{2^k} 种可能的模型可供选择。随着分类变量数目的增加，可选择的模型数目呈现爆炸式增长，从中寻找最优的模型变得非常困难。当分类变量数目大于 3 时，用所有的可能模型来拟合数据，并选取其中最优模型的方法是不现实的。

因此，人们提出了一些模型选择策略[7]，主要有三种策略，第一种策略是层次模型选择，首先选择所有单因子，并计算该模型的拟合度和自由度，然后选择所有二阶因子与单因子，计算该模型的拟合度和自由度，如此过程直到全饱和模型。在这个过程中，模型的拟合度和自由度都在逐渐变少，它能够严格地证明模型拟合度优劣。例如，与含有所有二

阶因子和所有单因子的模型相比，含有所有三阶因子及以下因子的模型具有更好的拟合度。因此，可以根据拟合度以及因子个数来选择合适的模型。

第二种策略在模型选择过程中，采用两个测试方法来考查所有因子的重要性，首先测试出简单模型的最复杂参数，然后再考虑其他所有的参数复杂度。这个策略过程需要评估所有因子的重要性，涉及到大量的计算过程，耗费时间较长。

第三种策略在模型选择过程中，考查系数 γ 与其标准误差 $\sigma(\gamma)$ 的比率，即比率 $\gamma/\sigma(\gamma)$，也称为标准化参数估计。标准化参数估计 $\gamma/\sigma(\gamma)$ 越大，则表明系数 γ 越显著。在一个拟合的模型中，为了使所有包含 γ 的系数都显著，则相应的所有标准化参数估计绝对值需要超过一定的阈值，研究表明，较理想的阈值是 2。

下面采取类似于第三种策略来选择模型，即当 γ 系数的所有标准化参数估计绝对值都低于阈值 2 时，则 γ 系数排除在拟合模型外。根据这一策略，可以得到一个精简的拟合模型。

5.2.3 微博转发行为特性分析

下面通过实验数据对微博用户转发行为特性进行分析。

1．实验数据集

实验数据来源于新浪微博。数据集采集于 2011 年 5 月至 7 月，随机选取了 3430 个种子用户以及其关注的 171769 个用户，然后排除不活跃的关注用户，数据集最终收集到 702789 条活跃关注边，其中 185327 条边含有转发记录。

不活跃的关注用户是指在关注边中存在不活跃的一些关注用户，他们几乎从不发布微博，很难有转发行为，因此需要从数据集中排除不活跃的关注用户。对于关注用户的活跃程度，采用下式来定义：

$$\theta = \frac{T}{R} \tag{5-7}$$

式中，T 和 R 分别表示发布微博的数目和关注用户注册时间长短，θ 表示关注用户活跃程度，θ 值越大，则该用户越活跃，反之亦然。对于关注者的活跃阈值，设定 $\theta=2$，即当 θ 大于 2，为活跃用户，否则为不活跃用户。

2．微博数据分布

表 5-3 给出了所有活跃关注边的微博转发数据分布，其中特征及相应的值如表 5-2 所示。在表 5-3 中，特征 A 对应于权威比率，相应索引值 1、2、3 分别代表 Small、Medium 和 Large，每个单元格的值对应于关注边 $A \to B$ 相应属性值的个数。例如，表 5-3（a）的最左上角单元格的值为 1059，则表示权威比率为 Small、微网络结构为 Pattern I、地理距离为 Near、性别为 MM 以及有转发记录的活跃关注边的数量是 1059 条。

表 5-3　（a）微博转发数据分布 I

		D	1						2					
		C	1		2		3		1		2		3	
A	B	E	1	0	1	0	1	0	1	0	1	0	1	0
1	1		1059	3563	1663	5804	3816	22441	771	3594	1391	5991	4463	19044
	2		35	333	49	636	364	3841	44	610	107	1132	447	5062
	3		254	2027	441	4305	1704	19840	186	3096	306	5122	1343	22679
2	1		621	982	1036	1735	2178	6765	145	473	316	996	760	3460
	2		32	75	51	189	323	1039	28	51	45	140	168	727
	3		1230	4093	2736	10917	12400	47840	435	2270	1392	5574	5708	24438
3	1		24	37	79	65	229	214	17	9	19	22	45	93
	2		1	4	4	9	31	34	6	3	1	3	11	24
	3		1525	2239	4722	6421	23209	31536	988	1226	3142	3623	19132	21533

表 5-3　（b）微博转发数据分布 II

		D	3						4					
		C	1		2		3		1		2		3	
A	B	E	1	0	1	0	1	0	1	0	1	0	1	0
1	1		590	1919	912	3404	2409	11772	1029	2548	1455	4284	4551	16147
	2		29	298	41	636	316	2844	28	320	67	718	304	3945
	3		80	809	187	1693	732	10374	112	1192	236	2644	1811	16417
2	1		241	525	469	1037	1260	4168	233	319	341	693	816	2661
	2		6	70	40	191	188	909	3	68	25	116	151	640
	3		463	2109	1563	5666	7800	27443	326	1411	980	3501	5854	19016
3	1		156	24	52	48	68	138	19	8	62	20	79	62
	2		0	2	3	7	33	50	0	3	0	5	24	24
	3		1142	1454	2772	4269	20647	25913	1135	862	2570	2929	19685	21428

图 5-3 给出了所有活跃关注边 $A \rightarrow B$ 中用户关注边数的分布，它符合幂律分布，在大部分关注边中两用户权威差值较小，而小部分关注边的两用户权威差值较大，这说明大部分关注边都是地位相称的用户。

图 5-4 给出了所有关注边 $A \rightarrow B$ 中用户转发次数的分布，它也符合幂律分布，大部分用户的微博几乎没有或者很少被转发，而小部分用户的微博则被大量转发，尤其存在少数用户转发数高达 100 次以上。由此可见，虽然用户关注很多其他用户，但是真正感兴趣的只是其中极少部分用户。

3．转发行为特性分析

下面讨论权威比率与微网络结构、地理距离与微网络结构、性别与微网络结构等两两组合属性对转发因子的权重，属性组合的权重越大，表示对转发的影响越大。

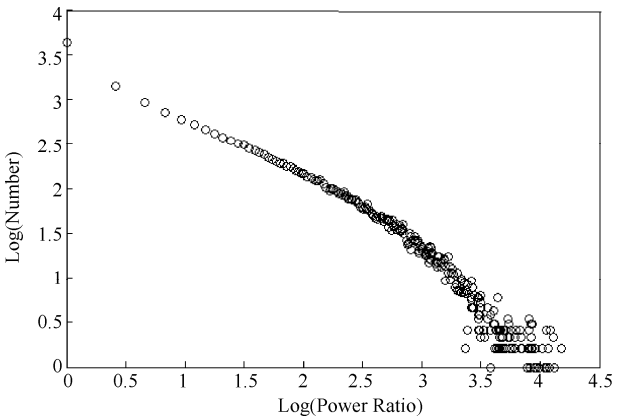


图 5-3　用户的关注边数分布

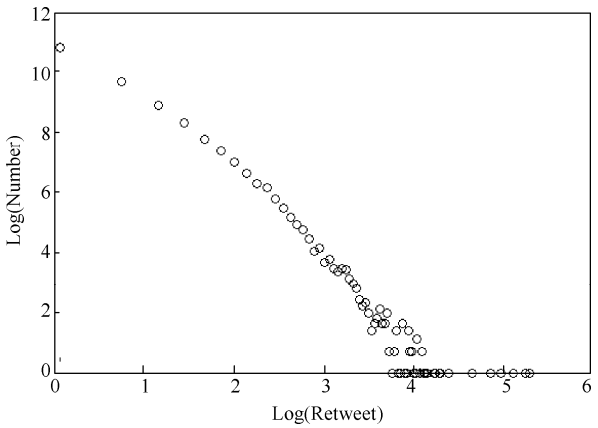


图 5-4　关注边微博转发次数分布

（1）权威比率、微网络结构对转发影响的权重。从表 5-4 可以看出，在 Pattern I 中，权威比率为 Small 的权重比较大，表明权威比率影响较小，说明在 Pattern I 中包含了大量的强连接关系，社会地位相当的用户所占比例更大一些。在 Pattern II 中，权威比率为 Medium 的权重比较大，表明权威比率影响较小，说明 Pattern II 中包含了大量的弱连接关系。在 Pattern III 中，权威比率为 Large 的权重最大，说明在 Pattern III 中包含了更多的权威连接关系。

（2）地理距离、微网络结构对转发影响的权重。从表 5-5 可以看出，在 Pattern I 中，地理距离为 Near 的权重比较大；在 Pattern II 中，地理距离为 Far 的权重比较大；在 Pattern III 中，地理距离为 Medium 的权重比较大。这说明 Pattern I 和 Pattern II 容易受到地理距离的影响，Pattern I 中的用户之间的地理距离更近，而 Pattern II 中的用户之间的地

理距离更远。这是因为在强连接关系中，用户为了维持这种关系需要花费大量的时间、情感等，而近距离比远距离更容易维持这种关系。在弱连接关系中，远距离的用户反而更容易被关注，其原因是远距离的用户更多来自不同的社区或团体，可能包含更多的新颖的信息，更容易受到用户的关注。在 Pattern III 中，大多数是权威连接关系，一般用户只需通过公共媒体，例如电视、新闻等，单方面地了解权威人士，因此几乎不受地理距离的限制。

表 5-4　权威比率、微网络结构对转发影响的权重

		微网络结构		
		Pattern I	Pattern II	Pattern III
权威比率	Small	0.078	0.063	–0.141
	Medium	–0.114	0.119	–0.005
	Large	0.036	–0.182	0.146

表 5-5　地理距离、微网络结构对转发影响的权重

		微网络结构		
		Pattern I	Pattern II	Pattern III
地理距离	Near	0.133	–0.128	–0.005
	Medium	–0.045	–0.051	0.096
	Far	–0.088	0.179	–0.091

（3）性别、微网络结构对转发影响的权重。从表 5-6 可以看出，在 Pattern I 中，无论男性用户还是女性用户的微博，都会受到女性用户的关注（FM、FF），这是因为在 Pattern I 中包含了大量的强连接关系，用户之间大多是朋友、亲戚关系，女性用户更愿意与同性朋友交往。在 Pattern II 中，无论男性用户还是女性用户的微博，都会受到男性用户的关注（MM、MF），这是因为在 Pattern II 中包含了大量的弱连接关系，用户一般来自不同的社区或团体，能够提供新颖的信息，男性用户比较关注新颖的信息。在 Pattern III 中，无论男性用户还是女性用户的微博，都会受到女性用户的关注（FM、FF），说明服从权威因素对女性的影响更大一些。

表 5-6　性别、微网络结构对转发影响的权重

		微网络结构		
		Pattern I	Pattern II	Pattern III
性别	MM	–0.080	0.111	–0.031
	MF	–0.141	0.215	–0.074
	FM	0.043	–0.062	0.019
	FF	0.178	–0.264	0.096

（4）权威比率对转发影响的权重。随着权威比率的增大，对微博转发影响的比重也在

增大，这意味着权威比率增大导致了微博更容易转发，表明权威比率对转发行为有着较大的影响，权威比率与转发行为之间有着很强的关联性。可以解释为随着权威比率的增大，社会权威差距在增大，服从权威效应逐渐显露出来，导致用户容易转发权威人士的信息。因此，服从权威对转发行为有着显著地影响。

（5）微网络结构对转发影响的权重。在 Pattern I 中，容易发生转发行为，这是因为 Pattern I 包含了强连接关系，两个用户大多来自于相同的社区或团体，有着共同的价值观和认同感，因此更加容易接受对方的信息。相反地，在 Pattern II 中，不容易发生转发行为，这是因为 Pattern II 中包含了弱连接关系，用户可能来自于不同的社区或团体，他们的价值观也不一样，因此较难接受对方的信息，除非是新颖的信息。在 Pattern III 中，与 Pattern I 中用户较容易转发和 Pattern II 中用户较难转发相比，Pattern III 中的用户转发行为处于中间水平，表明用户对权威连接关系的微博转发意愿处于强连接与弱连接中间。

综上所述，微网络结构、权威比率、地理距离及性别等 4 个属性对微博转发行为具有不同的影响：

（1）不同的微网络结构对微博转发行为的影响不同。Pattern I 中包含了较多的强连接关系，用户一般来自于相同的社区或团体，有着共同的价值观和认同感，更加容易接受对方的信息，用户之间更容易产生微博转发行为。Pattern II 中包含了较多的弱连接关系，用户一般来自于不同的社区或团体，他们的价值观也不一样，接受对方的信息比较困难，用户之间不容易产生微博转发行为。Pattern III 中包含了权威连接关系，用户对权威连接关系的微博转发意愿处于前两者之间。

（2）权威比率对微博转发行为的影响因微网络结构而异。权威比率对 Pattern I 中的微博转发行为影响较小，这是因为 Pattern I 中包含了较多的强连接关系，微博转发比较容易，而权威比率的增大并不会对微博转发行为产生更大的影响。权威比率的增大将使得 Pattern II 中的微博转发更加困难，这是因为来自于不同社区或团体的用户对权威人士或名人的抵触心理比较强烈，很难转发他们的微博。权威比率对 Pattern III 中的微博转发行为影响最大，随着权威比率增大，微博转发更加容易，服从权威效应逐渐显露出来，用户容易转发权威人士的信息。

（3）地理距离对微博转发行为影响与微网络结构有关。在 Pattern I 中，近距离用户的微博更容易转发，因为较近的地理距离分布着更多的强连接关系。在 Pattern II 中，远距离用户的微博更容易转发，因为远距离的用户微博包含更多的新颖信息，用户对新颖的微博怀有更多的好奇心。在 Pattern III 中，不同地理距离对转发影响不大，这表明权威连接关系的建立不受用户之间地理距离的影响。

（4）性别对微博转发行为影响与权威比率有关。女性用户比较关注具有强连接关系的朋友微博，更愿意与同性朋友交往。男性用户比较关注来自不同社区或团体的新颖信

息，更关注女性的微博。服从权威因素对女性的影响更大一些，说明女性更喜欢崇拜网络名人。

5.3　微博转发行为预测

随着 Web 2.0 的发展，微博网络已经成为互联网中最流行的信息共享和分发平台。微博转发是微博网络中最重要的信息传播机制，如果能够准确地预测微博用户转发行为，那么就能够预测该微博的传播方向、次数以及覆盖范围等，对于网络舆情发现、突发事件预测以及用户推荐等方面具有重要的现实意义，因此越来越受到研究者的重视。

下面给出一种基于社会纽带关系的微博转发预测方法，首先基于社会纽带关系提取用户之间的微网络结构、权威比率以及其他特征，然后根据这些特征对微博转发的权重大小，采用基于随机森林的预测算法对微博转发行为进行预测。

5.3.1　决策树算法

随机森林是一种分类预测算法[8]，由于决策树是随机森林的基本分类器，因此首先介绍决策树及相关概念。

决策树是一种类似流程图的树结构，由节点和有向边组成。树中包括三类节点：根节点、内部节点和叶子节点。其中，根节点位于树的最顶层；内部节点代表一个属性分裂问题，每个分裂输出代表一个分枝；叶子节点是终节点，存放着带分类标签的数据集。从根节点到叶子节点的每一条路径都形成一个分类。决策树的算法有多种，如 ID3、C4.5 以及 CART 等，这些算法通常采用自上而下的贪婪算法来构造，通用的决策树算法步骤如下。

算法 5-1　通用决策树算法

输入：　训练数据集 D
　　候选属性的集合 attribute_list
　　划分准则 Attribute_selection_method，由分裂属性和分裂点组成
输出：一棵决策树
（1）创建一个节点 N;
（2）If D 中元组都是同一类 C then
（3）　　返回 N 为叶子节点，以类 C 标记；
（4）If attribute_list 为空 then
（5）　　返回 N 为叶子节点，标记为 D 中的多数类；
（6）使用 Attribute_selection_method（D， attribute_list），找出最好 splitting_criterion；
（7）用 splitting_criterion 标记节点 N；

（8）If splitting_attribute 是离散的并且允许多路划分 then
（9） 从 attribute_list 中删除属性 splitting_attribute；
（10）For splitting_criterion 的每个输出 j
（11） 设 Dj 是 D 中满足输出 j 的数据元组的集合；
（12） If Dj 为空 then
（13） 加一个树叶到节点 N，标记为 D 中的多数类；
（14） Else
（15） 加一个由 Generate_decision_tree（Dj， attribute_list）返回的节点到节点 N;
（16）End for
（17）返回 N;

该算法需要输入三个参数：D、attribute_list 和 Attribute_selection_method。其中，D 是训练数据集，attribute_list 是描述数据元组的属性列表，Attribute_selection_method 是指定选择属性的启发式方法，该方法通常使用一种属性选择度量，其中包括信息增益、信息增益比率以及基尼（Gini）指数等。通用的决策树算法过程都基本类似，而各个算法的差别在于在创建树时属性选择方法以及剪枝方法。

1．属性选择度量

属性选择度量是针对属性分裂设定准则，将给定的训练数据划分成不同的子集。属性选择度量有很多种，下面主要介绍 ID3[9]、C4.5[10]以及 CART[11]算法中用到的度量方法。

ID3 算法是经典的决策树方法，用信息增益作为属性选择度量。若属性 α 的值将样本集 D 划分成 $D_1,D_2,\cdots,D_n$，共 n 个子集，那么信息增益定义如下：

$$\mathrm{Gain}(\alpha)=\mathrm{Entrop}(D)-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\times\mathrm{Entrop}(D) \tag{5-8}$$

式中，$|D|$为 D 的样本个数，$|D_i|$为子集 D_i 的样本个数，$\mathrm{Entrop}(D)$ 为信息熵，可以由下式得到：

$$\mathrm{Entrop}(D)=-\sum_{i=1}^{m}p_i\log_2(p_i) \tag{5-9}$$

式中，p_i 为 D 中样本属于类 C_i 的概率，由$|C_{i,D}|/|D|$估计得到。信息增益表示整体样本集合 D 的熵相对于用属性 α 对 D 分类得到 n 个子类后，各个子类的加权平均熵的差值，其中加权值为各子类 D_i 占全部样本的比例 D_i/D。

由于信息增益度量偏向于具有多输出的属性，ID3 的后继算法 C4.5 使用了信息增益比来克服这种偏向。

C4.5 算法的信息增益比定义如下：

$$\text{GainRation}(\alpha)=\frac{\text{Gain}(\alpha)}{\text{SplitInfo}(\alpha)} \tag{5-10}$$

式中，$\text{Gain}(\alpha)$ 如公式（5-8），分裂信息 $\text{SplitInfo}(\alpha)$ 定义如下：

$$\text{SplitInfo}(\alpha)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\times\log_2\left(\frac{|D_i|}{|D|}\right) \tag{5-11}$$

公式（5-11）表示样本集 D 被属性 α 划分 n 个子集而生成的信息。实际上 $\text{SplitInfo}(\alpha)$ 就是将每个样本视为等可能情况下的熵。

CART（Classification And Regression Tree）算法则是采用 Gini 指数作为尺度来分裂属性的，其中 Gini 指数用来定义样本集 D 的不纯度，即：

$$\text{Gini}(D)=1-\sum_{i=1}^{k}p_i^2 \tag{5-12}$$

式中，p_i 为目标变量不同类别在样本集 D 中的概率，用 D_i/D 估计。当 $\text{Gini}(\beta)=0$ 时，属性将所有样本划分一个类别，信息最大；当 $\text{Gini}(\beta)$ 最大，属性将样本均分在所有类别中，服从均匀分布，信息最少。Gini 指数适用于二进制、连续值等类型的字段，如果属性 β 将 D 二元划分为 D_1 和 D_2，则有：

$$\text{Gini}(\beta)=\frac{|D_1|}{|D|}\text{Gini}(D_1)+\frac{|D_2|}{|D|}\text{Gini}(D_2) \tag{5-13}$$

对于属性 β 的二元分裂导致不纯度下降值为：

$$\text{Gini}(\beta)=\text{Gini}(D)-\text{Gini}(\beta) \tag{5-14}$$

CART 算法选择具有最大 $\text{Gini}(\beta)$ 的属性作为分裂属性。ID3（C4.5）算法选择准则是计算每一个候选属性的信息增益（信息增益比），然后选择信息增益（信息增益比）最大的属性进行分裂。

上述三种属性选择度量方法各有偏向，信息增益偏向于多值属性，尽管信息增益比调整了这种偏向，但是它倾向于不平衡的分裂，其中一个划分比其他划分小得多。Gini 指数也偏向于多值属性，并且当属性的类很大时会有一定的困难，并且它也倾向于导致大小相等的划分和纯度。尽管这些方法是有偏向的，但是它们在实践中取得较好的效果。

另外，还有一些其他的属性选择度量方法，例如，CHAID 算法[12]使用一种基于独立统计卡方检验的属性选择度量方法。

2．决策树剪枝

在决策树创建时，由于数据中的噪声和离群点，许多分枝反映的是训练数据的异常，

需要剪去这些不可靠的分枝。剪枝方法用来处理这种过分拟合问题，通常有先剪枝和后剪枝两种方法。

在先剪枝方法中，通过提前停止树的构造来实现对树的剪枝。例如，在构造树的时候，可以使用诸如统计显著性、信息增益、Gini 指数等度量来评估分裂的优劣。如果划分一个节点的属性值低于预定义的阈值，则停止分裂。

后剪枝方法中，首先让树充分生长之后，再判断是否剪去一些分枝。常用的方法包括根据错误分类率对决策树进行事后修剪等。

在决策树学习过程中，虽然剪枝可以减少过度拟合问题，但是剪枝方法以及相应的度量选取最终影响决策树的优劣，而这一选取过程是决策树学习过程中的一个难点。另外，单棵决策树的准确度也容易受到训练数据的影响而导致分类结果的不稳定。为了克服这些缺点，引入一个新的预测模型，即随机森林模型。

5.3.2 随机森林算法

单棵决策树对数据分类有较好的准确度，然而这种分类准确度易受到训练数据本身的影响，具有不稳定性。为了避免这种不稳定性，一种解决方案是产生多棵决策树，参与投票，选出最好的分类，这就是随机森林的思想。

随机森林就是采用随机的方式建立一个森林，森林里面包括多棵决策树，且每一棵决策树之间没有关联。在建立森林后，当有新的输入样本数据进入时，就让每棵决策树进行判断分类，当所有树判断分类结束，组合多棵树的预测，通过某种投票方式得到最终预测结果。

1．算法构造

随机森林算法定义为由一组决策树分类器 $(h(X,\theta_K),k=1\cdots K)$ 组成的集成分类器，其中 $\{\theta_k\}$ 是服从独立同分布的随机变量，K 表示随机森林中决策树的个数，在给定自变量 X 下，每个决策树分类器通过投票来决定最优的分类结果。

随机森林算法一般构造过程如下。

（1）给定全部训练数据，随机抽取部分数据，形成新的子样本数据；

（2）对新的子样本数据中 M 个特征变量，随机抽取 m（m<M）个特征，然后构造完整的决策树；

（3）重复步骤（1）、（2），得到 K 个决策树，形成随机森林；

（4）每个决策树参与投票，最终以某种投票方式，选出最优的分类。

图 5-5 给出了通用随机森林构造过程图，从图 5-5 可以看出，所有决策树分类器都是并行的，也就是说每个决策树分类器的构造过程互不影响，相互独立。

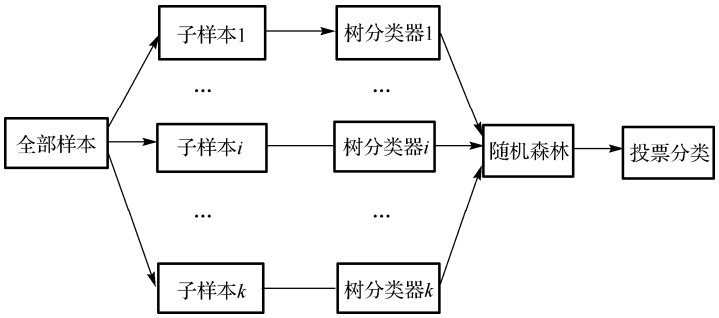


图 5-5　随机森林构建过程图

对于子样本采集过程，随机森林采用了两个随机采样方法，保证了后续的决策树分类器之间的独立性。

首先，该算法对输入的数据进行采样，即行采样。在此过程中，一个常用的方法是 bootstrap 方法，它是一种有放回的抽样方法，也就是在抽样得到的样本集合中，允许重复抽样。假设输入样本为 N 个，那么重复抽样的样本也为 N 个，这样使得在训练时抽取的子样本并不是全部的样本，当 N 值足够大时，约 63.2%输入样本成为子样本。数据采样还有其他采样方法，如保持、K 折交叉验证等方法。

其次，该算法对采样子数据的特征再次进行采样，即列采样。该过程通常从 M 个特征中以某种随机方法选取 m 个（$m \ll M$）特征，关于 m 的取值下面将详细地介绍。列采样通常有两种方法，第一种方法是随机选择特征，每一个节点随机选取一小组输入变量进行分割，节点分割是根据 F 个选定特征，而不是所有的特征来决定，然后利用 CART 方法完全生成决策树，一旦决策树完全构建完成，使用多数表决法来预测，在构建过程中选择的输入变量 F 是固定的，F 的大小影响随机森林的强度和相关性，如果 F 足够小，树的相关性减弱，同时分类模型的强度也在减弱。由于每个节点只需要考查输入变量的一个子集，该方法的运行时间较少。第二种方法是随机选取特征变量的线性组合，随机选择 L 个输入变量进行线性组合得到新的特征，在每个节点上，随机选出 L 个变量 $v_1, v_2, \cdots, v_L$ 以及相应的 L 个随机数 $k_1, k_2, \cdots, k_L$，且每个随机数 $k_i \in [-1,1]$，做线性组合 $V = \sum_{i=1}^{L} k_i v_i$。

与一般决策树构造不同的是，随机森林决策树不需要剪枝，即所有的决策树都是完全分裂的。在这个过程中，所有决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本都是指向同一个分类。一般通用的决策树算法都需要进行剪枝，但是随机森林并不需要这样做，一个重要的原因是两个随机的采样过程保证了数据的随机性和独立性，不会出现 over-fitting 现象。

另外，不同随机森林的投票机制也不同，当一个被预测实例进入，每个决策树分类器

都要进行预测分类，然后以某种方式参与投票。因此，不同的投票机制将导致不同的分类结果。投票机制主要分为两大类：简单投票机制和贝叶斯投票机制。

简单投票机制的基本思想是多个基本分类器都进行分类预测，然后根据分类结果用一种投票方法进行表决。投票方法可以分为一票否决、少数服从多数和阈值表决等方法。一票否决法是指预测结果当且仅当所有决策树分类器都把预测的实例划分到某一类时才有效，否则拒绝这个实例分类。少数服从多数法是指每个决策树分类器预测分类，然后进行投票，得票多数的类作为实例的最终分类结果。阈值表决法是指统计实例被决策树分类器划分和不划分某类的得票数，当两者比例超过预定义的阈值时，就划分到此类中。

贝叶斯投票机制不同于简单投票机制，简单投票机制假设每个决策树都是平等的，没有分类能力的差别，但是这种假设并不总是合适的。例如，在实际生活中，听取一个人的意见时会考虑这个人过去的意见是否有用。贝叶斯投票方法就是基于这种思想提出的，贝叶斯投票方法是基于每个决策树在过去分类的表现来设定一个权值的，然后按照这个权值进行投票，其中每个决策树权值可以利用贝叶斯定理计算出来。

理论上，贝叶斯投票方法在假设空间中所有假设的先验概率都正确的情况下能够获得很好的效果，但是在实际应用中往往不可能穷尽整个假设空间，也不可能准确地给每个假设分配先验概率。因此，在实际使用过程中，简单投票方法优于贝叶斯投票方法。

2．参数优化

决策树分类器个数 K 以及特征个数 m 的选取也直接影响着随机森林的性能，下面讨论如何选取最优参数 K 和 m。

随着决策树分类器个数 K 的增加，随机森林不会出现过度拟合问题，但是会产生一个有限的泛化误差。分类器泛化误差定义为：

$$PE = P_{X,Y}(mg(X,Y) < 0) \tag{5-15}$$

式中，PE 为泛化误差，$P_{X,Y}$ 的下标 X,Y 表示概率覆盖的定义空间。

分类器泛化误差的上界为：

$$PE \leqslant \bar{\rho}(1-s^2)/s^2 \tag{5-16}$$

式中，$\bar{\rho}$ 为分类器相关性均值，s 为分类器效能强度。

可见，随机森林的泛化误差上界包含两个要素：决策树的分类效能强度 $\bar{\rho}$ 和决策树间的相关性 s，可以由 $\bar{\rho}$ 和 s 两个参数推导出来，其中与 $\bar{\rho}$ 成正比，与 s^2 成反比，因此 $\bar{\rho}/s^2$ 比值越小越好。对于多于两个分类的情况，效能强度依赖于随机森林以及相应的每个决策树。

随着特征个数 m 的增加，分类器的相关性 $\bar{\rho}$ 和效能强度 s^2 也相应地增加，反之亦

然。因此，m 值会产生一定的泛化误差，当 m 值在某一区间时，泛化误差上界将处于最低。

3．特征提取

在分类预测中，一个重要的任务是寻找相关的重要特征。一方面在数据集中有许多特征与分类预测无关，另一方面有些特征可能是冗余的。如果选择的特征不具有辨别能力，则会直接影响到分类器性能。如果选择了具有充分辨别能力的特征，则会极大地提高分类器的预测精度。因此，特征选取过程是至关重要的。

微博转发行为在社交网络信息传播中有两个重要过程：同化与社会影响，同化过程导致了用户之间网络结构的变化，而社会影响过程则导致了两个用户的属性变化。用户之间不同类型的网络结构以及属性关系代表着不同类型的用户关系，这意味着微博是否会存在转发行为，这些特征对预测微博转发行为具有重要的作用。因此，需要从网络结构和用户属性中提取特征，包括权威比率、微网络结构、地理距离、用户性别以及用户自身属性值等。关于权威比率、微网络结构、地理距离、用户性别等特征参见 5.2.1 节，下面是用户属性特征的提取。

所谓微博转发行为预测是指对于任何关注边 $u_1 \to u_2$，预测用户 u_1 是否会转发用户 u_2 的微博，即预测在关注边 $u_1 \to u_2$ 上的微博转发。这里需要定义关注边概念，对于已知微博网络 $G=(V,E)$，V 为用户的集合，E 为有向关注边的集合，如果用户 $u_1 \in V$，$u_2 \in V$，且 u_1 关注 u_2，则关注边定义为：$u_1 \to u_2$。

关注边 $u_1 \to u_2$ 只是定义了用户 u_1 与其关注者用户 u_2 的关系，而没有给出用户 u_1 与被关注者 u_3 的关系，这种关系可以通过转化来生成相应的关注边。例如，用户 u_1 与被关注用户 u_3 的边 $u_1 \leftarrow u_3$，可以等价地转化为用户 u_3 关注用户 u_1，即 $u_3 \to u_1$。因此，微博中的所有关系都可以看作是关注边的关系。

对于关注边 $A \to B$，用户 A、B 都有基本资料，可以从中提取用户的自身属性特征，其中包括 9 个特征，如表 5-7 所示。用户 A、B 都有自身属性特征，因此共提取了 22 个相关特征。

表 5-7　用户自身属性特征

特征	描述
1	居住省份
2	居住城市
3	关注人数
4	粉丝数
5	微博数
6	喜欢的人数目

续表

特征	描述
7	注册时间
8	访问权限
9	是否为认证用户

4．算法实现

在随机森林预测算法中，一个重要的问题是数据随机采样。构造一组相互独立的决策树，数据采样包括数据集采样和数据特征采样。

对于数据集随机采样，经典的随机森林方法采用了 bootstrap 方法，即从数据集中有放回均匀抽样。在平均情况下，63.2%的数据集将作为样本子集训练模型，其余 36.8%数据集作为袋外测试数据集来检验模型。对于小数据集，bootstrap 方法效果很好。随着数据集增大，bootstrap 方法效果并不理想。由于微博数据集比较大，因此采用 K 折交叉验证方法来随机采集微博数据集。在该方法中，数据集被随机地采样，并划分为 K 个互不相交且大小一致的子集，然后利用数据对算法进行 K 次训练与测试，其中在第 i 次时，第 i 个子集作为测试数据，而将其余的所有数据子集一起作为训练数据。通常实验将采用 10 折交叉验证，这是因为它具有相对较低的偏向和方差。

对于数据特征的采样，经典的随机森林算法采用完全随机方法，即所有特征被抽取到的概率都相同。这里采用一种基于特征权重的方法来抽取特征，微博转发影响权重越大的特征越容易被抽取到，这是因为权重越大的特征对微博转发预测作用也越大，这样可以提高预测模型的准确度。其中，特征权重采用信息增益算法来刻画，通过特征的信息增益值来代表其权重大少，当特征的信息增益值越大，则该特征的影响权重也越大。

对于数据特征的采样，一方面，算法首先利用信息增益算法计算出所有特征的信息增益值并进行排序，然后根据特征的信息增益值去除对微博转发影响权重很小的特征，如果选取了那些权重很小的特征，则生成的决策树分类器区分度很弱，反而导致预测误差增大。另一方面，算法将根据不同特征的权重来选取特征，对微博转发影响权重越大的特征，被抽取到的概率也越大，有利于提高决策树分类器的准确度。

对于决策树的特征属性度量方法，采用 Gini 指数，即 CART 决策树。这是基于两方面考虑，一方面是部分选取的特征是连续的，CART 决策树能够处理这类特征；另一方面是从微博数据集中所提取的数据特征类的数量较少，CART 决策树能够很好地处理这类数据，不会因类的数量太少而产生度量偏差的问题。

对于分类预测的投票机制，采用简单多数投票法，即数目最多的类就是最终的类，分类决策公式如下：

$$H(x)=\arg\max_{Y}\sum_{i}I(h_i(x)=Y) \tag{5-17}$$

式中，$H(x)$ 表示组合分类模型，$h(x)$ 表示单个决策树分类模型。

一种改进的随机森林微博转发预测算法（IRFMR）具体步骤如下。

算法 5-2　随机森林微博转发预测算法（IRFMR）

输入：微博数据集 S

　　微博预测数据集 P

模型训练：

（1）对数据集 S 用 10 折交叉验证方法采样，得到新的训练数据集 S_n；

（2）对数据集 S_n，计算信息增益算法计算每个特征的权重，排序并排除小于设定阈值的特征；

（3）对于训练集 S_n 所有大于设定阈值的 M 个特征，基于特征的权重大小，随机选取 m（$m<<M$）个特征，构成新的数据集 S_m；

（4）对数据集构造完整的决策树（CART 方法），不进行剪枝；

（5）循环步骤（1）、（2）、（3）、（4），直到 K 个决策树建立，随机森林构造完成。

预测：

（6）对数据集 P 的一个变量 x 分类标签，每棵决策树进行投票；

（7）计算所有投票数 H（x），票数最高的分类就是变量 x 的分类标签；

（8）循环（6）、（7），直到数据集 P 所有变量的分类标签被标记。

输出：预测数据集 P 的分类标签。

5.3.3　算法验证

下面通过实验数据对微博转发行为预测算法性能进行测试和验证。

1．实验数据集

实验数据来源于新浪微博。从 2011 年 5 月至 2011 年 7 月，随机采集了 171769 个用户以及相应的 702789 条活跃的关注边，其中用户包括标签、个人资料等信息，关注边则包括两个用户转发关系。如果一条关注边中两个用户存在转发行为，则该关注边为正例，否则为负例，最终得到 185237 个正例和 517552 个负例。

Weka 是一个公共的数据挖掘与分析平台，集合了大量数据挖掘算法，包括了数据预处理、分类、回归、聚类、关联规则等，因此所有实验数据都在 Weka 平台上运行。

2．参数优化

算法对全部数据进行采样，有部分数据不在训练样本中，这些数据称为袋外数据（OOB），使用袋外数据测试模型性能称为 OOB 误差估计。在随机森林算法中，对每一棵

决策树进行测试，可以得到一个 OOB 误差估计。对所有决策树的 OOB 误差估计取平均值，即可得到随机森林的泛化误差估计，实验证明，随机森林的 OOB 误差估计是无偏估计。

用 OOB 误差估计来选取决策树的个数 K 以及特征的个数 m，当 OOB 误差估计最小时，则参数 K 和 m 为最优。由于 K 和 m 两个参数都对 OOB 误差估计有影响，两个参数组合，需要进行 $K\times m$ 次 OOB 误差估计比较。为了简化实验过程，在估计一个参数时需要固定另一个参数值，这样只需比较 $K+m$ 次。另外，由于数据集较大，每次计算 OOB 误差估计值时间较长，抽样 5%数据集进行试验。

图 5-6 给出了当 $m=2$ 时决策树个数 K 与 OOB 误差估计值的曲线图，随着 K 值增加，OOB 误差估计值在减少，但是下降趋势不同。当 K 值处于 5～17 区间时，开始阶段 OOB 误差估计值下降较快，随后逐步地减弱。当 K 处于 17～20 区间时，OOB 误差估计值下降趋势明显地减弱，已趋于平稳，这说明 OOB 误差估计接近收敛。

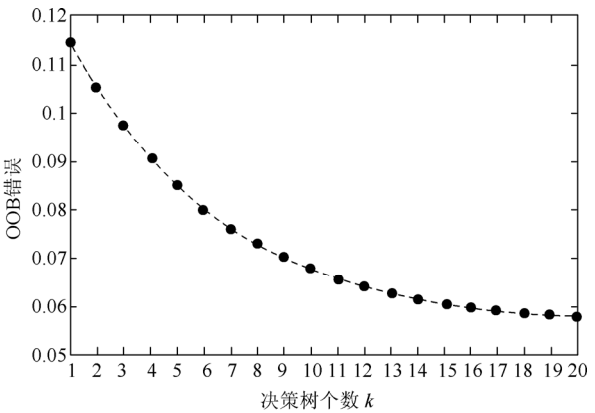


图 5-6　决策树个数 K 与 OOB 误差估计值曲线

图 5-7 给出了算法运行时间与 K 值的关系，从图 5-7 可以看出，随着 K 值增大，算法运行时间在逐步增加，综合考虑 OOB 误差估计及算法运行时间因素，则 $K=15$ 为最优。

图 5-8 给出了当 $K=15$ 时特征个数 m 与 OOB 误差估计的曲线变化，从图 5-8 可以看出，随着 m 值增加，OOB 误差估计值先迅速地下降，然后逐步地回升，当 $m=3$ 时，它处于最低值，实验结果最优。因此，将 IRFMR 算法的参数设定为 $K=15$，$m=3$。

3．算法性能对比

下面将 IRFMR 算法与逻辑回归（LR）、决策树（DT）、Adaboost（Ada）、朴素贝叶斯（NB）、多层感知器（MP）及经典随机森林方法（RF）等几种经典的分类算法进行对比实验。所有的算法都是在参数最优情况下得到的结果。例如，在经典随机森林算法中，

当决策树个数 $K=15$ 和特征个数 $m=3$ 时，该算法性能最优，选取此时的结果做比较。评价指标为准确率（P）、召回率（R）、F-度量（F_1）、ROC（Receiver Operating Characteristic），其中准确率、召回率、F_1 等指标的概念和计算公式参见 4.5.3 节。

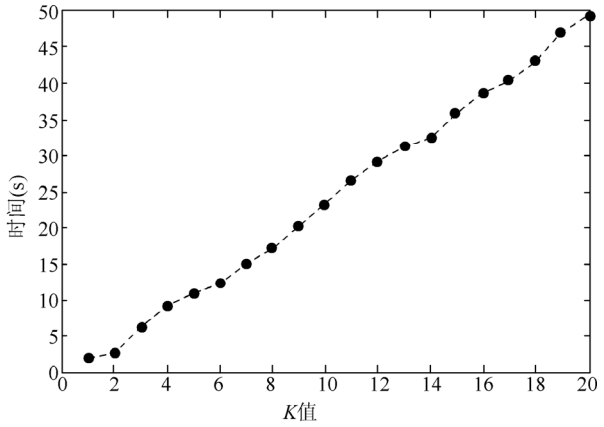


图 5-7　算法运行时间与 K 值的关系

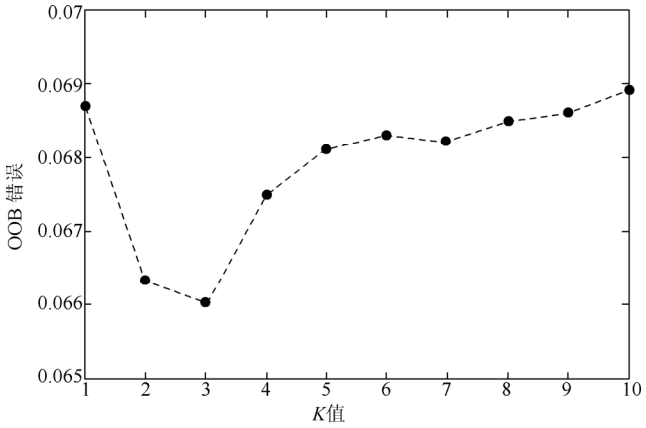


图 5-8　OOB 误差估计与 K 值变化曲线

表 5-8 给出了正例预测中各个算法比较结果，在正例预测中，不同算法效果相差明显。IRFMR 算法在各项指标上都明显优于其他算法，其中准确率为 94.8%，召回率为 70.3%，F_1 值为 80.7%，ROC 值为 94.9%，其次是 RF 算法，NB 算法表现最差。在召回率指标上，各个算法都不高，其原因是由于正负例的比例不均衡，大量被错误分类的负例个数降低了召回率。

表 5-9 给出了负例预测中各个算法的比较结果，相比于正例，各个算法的各项指标都较高，准确率和召回率都接近或超过 90%，说明所有算法在预测负例时效果都比较理

想。而 IRFMR 算法在各项指标上都优于其他算法，说明在负例预测中，IRFMR 算法性能也优于其他算法。

表 5-8　正例预测中各个算法性能对比

	P	R	F_1	ROC
LR	0.645	0.133	0.221	0.79
NB	0.359	0.282	0.316	0.74
DT	0.714	0.629	0.669	0.852
MP	0.709	0.254	0.374	0.816
Ada	0.615	0.142	0.231	0.794
RF	0.929	0.681	0.785	0.917
IRFMF	0.948	0.703	0.807	0.949

表 5-9　负例预测中各个算法性能对比

	P	R	F1	ROC
LR	0.883	0.989	0.933	0.79
NB	0.895	0.924	0.909	0.74
DT	0.945	0.962	0.953	0.852
MP	0.897	0.984	0.939	0.816
Ada	0.883	0.986	0.932	0.794
RF	0.953	0.992	0.977	0.917
IRFMF	0.957	0.994	0.979	0.949

综合所有的正负例预测，IRFMR 算法的准确率高达 95%，表明该算法可以成功地预测 95%的微博转发。因此，IRFMR 算法能够很好地预测用户间微博转发行为。

4．特征分析

采用信息增益算法来分析所提取的特征对微博转发预测的作用，特征的信息增益值越大，则该特征在微博转发预测的作用也就越大。图 5-9 给出 22 个特征值的权重值，其中 NS、GE、PR、LC 分别表示用户间的微网络结构、性别关系、权威比率、地理位置，A_1～A_9 分别表示用户 A 的 9 个自身属性特征，而 B_1～B_9 表示用户 B 的 9 个属性特征，与用户 A 属性特征相同，用户自身属性特征如表 5-7 所示。

从图 5-9 可以看出，权威比率的系数最显著，是其他特征值 3 倍以上，这说明权威比率在预测微博转发的作用最大，其次分别是微网络结构、性别关系以及地理位置。而用户 A、B 自身属性特征的系数总体上偏弱，尤其一些特征系数已经接近零值，这说明用户自

身属性特征对微博预测的作用较小。从以上分析可知，GE、LC、PR 和 NS 是微博转发预测中最重要的 4 个特征。

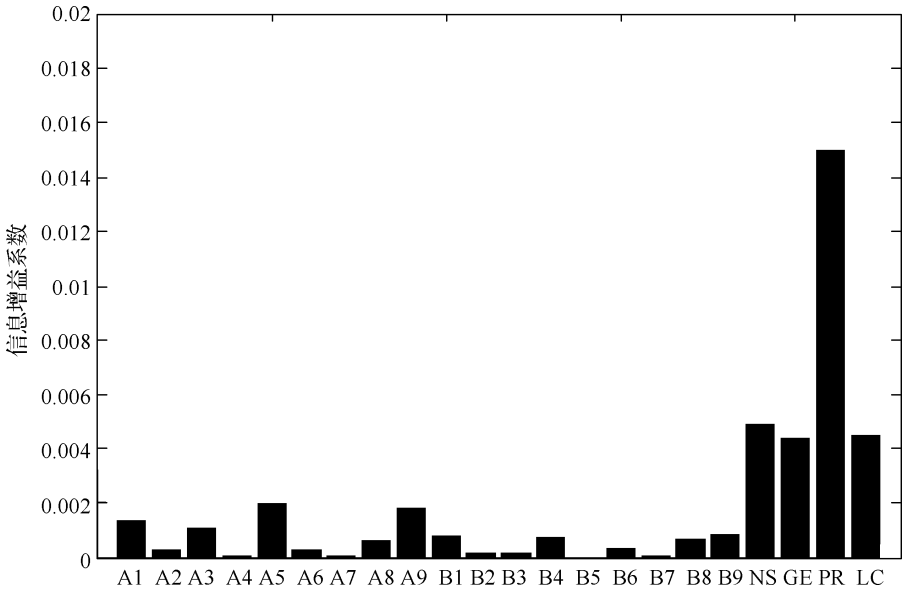


图 5-9　22 个特征值的权重值

综上所述，在微博网络中，权威比率、微网络结构、性别关系及地理位置等特征对微博用户转发行为产生重要的影响，通过提取这些特征，并采用 IRFMR 算法对微博用户转发行为进行预测，能够较好地预测微博用户转发行为。

5.4　微博转发特性预测

在微博网络中，用户发布一条微博后，如果微博内容有趣，该微博可能被关注者不断地转发，这样就使得一条微博在网络中迅速地传播，其传播速度以及范围呈现几何增长。通常，一条微博被转发次数越多、覆盖范围越大，则说明该微博影响力就越大。因此，微博转发次数和覆盖范围等转发特性是评价微博影响力的重要指标。

5.4.1　预测模型

当用户发布一条微博后，如何预测该微博在一定时间内的转发次数以及覆盖范围是研究微博网络信息传播机制和特性的重要问题，所谓覆盖范围是指该条微博发布后能够观察到该微博的用户总数。下面给出微博转发次数和覆盖范围预测问题的定义。

已知一条原始微博前 100 条转发信息，预测在 30 天内该条微博转发次数，则该预测定义为微博转发次数预测问题，简称 M1。

已知一条原始微博前 100 条转发信息，预测在 30 天内可观察该条微博的用户总数，则此预测定义为微博覆盖范围预测问题，简称 M2。

显然，用户微博的 M1 与 M2 存在着关联性。图 5-10 给出了某一时刻微博转发状态图，用户 Alice 发布的一条微博在微博网络中被 Bob、Cathy 以及 David 等用户所转发，并可以被 Ellen、Fred 等用户所观察到。该网络是一个有向的用户关注网络图，箭头表示用户关注方向。在该网络中，用户通过关注边观察到某一用户的微博，通过转发机制实现了微博在网络中的传播。

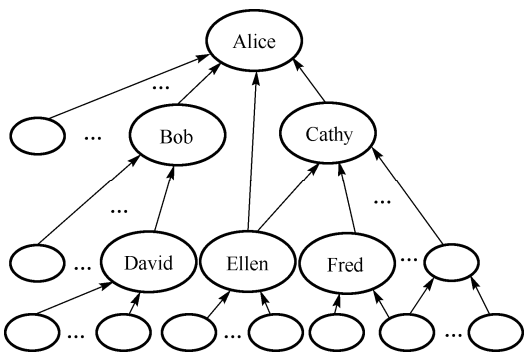


图 5-10　微博转发状态图

因此，微博在网络中转发次数（M1）越多，则观察到该条微博的用户也就越多，即覆盖范围（M2）越大。同时，随着微博 M2 值的增大，微博转发几率也随之增大，从而导致 M1 值也增大，M1 值与 M2 值相互促进，最终致使微博在整个网络传播。

用户微博转发次数与覆盖范围意义是不同的，各有侧重。微博转发次数是指已经转发该微博的用户数目，即已经接受该条微博的用户数量，它倾向于对一条微博影响力的评价，而微博覆盖范围则是指网络中观察到该微博的用户数目，这些用户不一定都转发，它更倾向于对传播范围大小的评价。虽然用户微博的转发数与覆盖范围可以相互促进，但是增长幅度并非完全同步。例如，一般用户多次转发的覆盖范围远不及超级用户一次转发的覆盖范围。因此，需要对微博的转发次数和覆盖范围分别进行预测，这样才能更好地理解微博影响力和传播范围。

虽然微博网络结构对微博转发具有重要的影响，但基于网络结构来预测微博转发次数与覆盖范围存在两个问题，一是静态的微博网络结构很难预测出动态的微博转发次数和覆盖范围，例如，某一个用户的网络结构比较稳定，但是他发布的微博可能被不同用户所转发，使得转发次数与覆盖范围有很大的不同，因此使用静态的网络结构很难准确地预测该

类微博的转发次数和覆盖范围；二是微博网络结构只是反映用户间的关注关系，它很难准确地反映用户间的转发关系，例如，某一个用户与很多用户建立了关注边，在网络结构中这些关注边的表现都是相同的，但是这些关注边的用户转发意愿是不一样的，一些用户转发微博比较积极，而另一些用户转发微博并不积极，单一的关注边并不能捕捉到用户转发意愿强度。

在社会媒体网络中，信息传播具有很强的时效性，信息传播在短时间内迅速地升至峰值，然后随着时间推移逐渐地呈指数衰减直至最终消失，其时间序列曲线特征非常明显。微博作为一种特殊的社会媒体网络，其转发行为也具有相似性。因此，可以采用微博转发时间序列曲线来预测微博转发次数和覆盖范围。

基于时间序列的预测模型是要找出一个近似产生时间序列历史模式的数学公式，并用该公式对未来值做长期或短期的预测。该类预测模型的典型代表是 ARMA（Auto-Regressive and Moving Average）模型，又称为 Box-Jenkins 方法[13]，它提供了一套完整、正规和结构化的建模方法，包括对时间序列的分析、预测以及对 ARMA 模型的识别、估计和诊断等，并且具有统计学的理论基础。

ARMA 模型选择与时间序列的平稳性有着密切的联系，在介绍 ARMA 模型之前，首先介绍时间序列的平稳性。

设时间序列$\{y_t\}$，对于任意的 t，k 和 m，满足下式，则称$\{y_t\}$是平稳的：

$$\begin{aligned} E(y_t) &= E(y_{t+m}) \\ \operatorname{cov}(y_t, y_{t+k}) &= \operatorname{cov}(y_{t+m}, y_{t+m+k}) \end{aligned} \tag{5-18}$$

时间序列$\{y_t\}$取自某一个随机过程，如果该随机过程的随机特征不随时间变化，则称该过程是平稳的；如果该随机过程的随机特征随时间变化，则称该过程是非平稳的。

ARMA 模型有三种基本形式：AR（Auto-Regressive）模型、MA（Moving-Average）模型和 ARMA（Auto-Regressive Moving-Average）模型，它们是针对一个平稳的时间序列。对于非平稳的时间序列，通常采用扩展的 ARIMA 模型。

1．AR 模型

如果平稳时间序列 y_t 是它的历史值和随机项的线性函数，可表示为:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t \tag{5-19}$$

式中，随机项 e_t 为相互独立的白噪声序列，与滞后变量 $y_{t-1}, y_{t-2}, \cdots, y_{t-p}$ 不相关，且服从均值为 $E(e_t)=0$、方差为 $\sigma_e^2 > 0$ 的正态分布，则称该时间序列 y_t 服从 P 阶的自回归模型，记为 AR(p)。实参数 $\phi_1, \phi_2, \cdots, \phi_p$ 称为自回归系数，是模型待估参数。

记 B^k 为 k 步滞后算子，即 $B^k y_t = y_{t-k}$，则公式（5-19）可表示为:

$$y_t = \phi_1 B y_t + \phi_2 B^2 y_t + \cdots + \phi_p B^p y_t + e_t \tag{5-20}$$

引入滞后多项式，令 $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$，则可简化为：

$$\phi(B) y_t = e_t \tag{5-21}$$

AR(p)过程平稳的条件是滞后多项式 $\phi(B)$ 的根均在单位圆外，即 $\phi(B) = 0$ 的根大于 1。

2．MA 模型

如果平稳时间序列 y_t 是它的当前以及历史的随机误差项的线性函数，即可表示为：

$$y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{5-22}$$

则称该时间序列 y_t 是移动平均序列，公式（5-22）称为 q 阶移动平均模型，记为 MA(q)模型。实参数 $\theta_1, \theta_2, \cdots, \theta_q$ 为移动平均系数，是模型的待估系数。引入滞后算子 $\theta(B)$，并令 $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$，公式（5-22）可简化为：

$$y_t = \theta(B) e_t \tag{5-23}$$

在移动平均过程中，MA(q)模型总是平稳的。但是，经常需要将 AR 模型表示为 MA 模型，反过来也一样，即具有可逆性。

对于 MA 模型的可逆性，则要求滞后多项式 $\theta(B)$ 的根都在单位圆外，经过推导可得：

$$(1 - \pi_1 B - \pi_2 B^2 - \cdots) y_t - \left(-\sum_{j=0}^{\infty} \pi_j B^j \right) y_t = e_t \tag{5-24}$$

式中，$\pi_0 = -1, B^0 = 1$，其他权重 π_j 可递推得到。公式（5-24）称为 MA(q)模型的逆转形式，它等价于无穷阶的 AR 过程。

3．ARMA 模型

如果平稳时间序列 y_t 是它的当期和历史的随机误差项以及历史值的线性函数，可表示为：

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t - \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \tag{5-25}$$

则该时间序列 y_t 称为自回归平均序列，公式（5-25）称为(p，q)阶的自回归移动平均模型，记为 ARMA(p，q)。$\phi_1, \phi_2, \cdots, \phi_p$ 为自回归系数，$\theta_1, \theta_2, \cdots, \theta_q$ 为移动平均系数，都是模型的待估参数。引入滞后算子 B，公式（5-25）可简化为：

$$\phi(B) y_t = \theta(B) e_t \tag{5-26}$$

ARMA(p, q)过程的平稳条件是滞后多项式 $\phi(B)$ 的根均在单位圆外，可逆条件是 $\theta(B)$

的根都在单位圆外。若 $\phi(B)=1$，则 ARMA(p, q)过程退化为 MA(q)过程，若 $\theta(B)=1$，则 ARMA(p, q)过程退化为 AR(q)过程。

4．ARIMA 模型

ARIMA 模型是 ARMA 模型的扩展。AR、MA、ARMA 模型都是针对平稳的时间序列进行建模的。而在实际情况中，不平稳的时间序列更为常见，必须通过差分后转化为平稳的时间序列，才能使用 ARIMA 模型。

对于不平稳时间序列 y_t，经过 d 次差分之后变换成平稳的 $\Delta^d y_t$，且 $\Delta^d y_t$ 是它的当期和历史的随机误差项以及历史值的线性函数，可表示为：

$$\Delta^d y_t=\phi_1\Delta^d y_{t-1}+\phi_2\Delta^d y_{t-2}+\cdots+\phi_p\Delta^d y_{t-p}+e_t-\theta_1 e_{t-1}+\theta_2 e_{t-2}+\cdots+\theta_q e_{t-q} \tag{5-27}$$

则该时间序列 y_t 称为自回归平均序列，公式（5-27）称为（p, k, q）阶的差分自回归移动平均模型，记为 ARIMA(p, k, q)。引入滞后算子 B，则公式（5-27）可简化为：

$$\phi(B)\Delta^d y_t=\theta(B)e_t \tag{5-28}$$

ARIMA(p, k, q)过程的平稳条件是滞后多项式 $\phi(B)$ 的根均在单位圆外，可逆条件是 $\theta(B)$ 的根都在单位圆外。若 $\phi(B)=1$，则 ARIMA(p, k, q)过程退化为 MA(q)过程，若 $\theta(B)=1$，则 ARIMA(p, k, q)过程退化为 AR(q)过程。由此可见，ARIMA 模型是 ARMA 模型的扩展。

5．模型选择

对于一个随机的时间序列，模型选择通常需要借助于分析工具，自相关（ACF）系数和偏自相关（PACF）系数是分析时间序列和识别模型的有力工具。根据绘制的自相关分析图和偏自相关分析图，可以初步识别出平稳序列的模型类型和模型阶数。利用自相关分析法可以测定时间序列的随机性和平稳性以及时间序列的季节性。

设已有平稳时间序列为 $X=\{x_1,x_2,\cdots,x_n\}$，期望值 $E(X)=\mu$，方差 $\mathrm{Var}(X)=\sigma^2$。相隔 k 期的两个随机变量 x_i 与 x_{i-k} 滞后 k 期的自协方差为 $\rho_k=\mathrm{Cov}(x_i,x_{i-k})=E[(x_i-\mu)(x_{i-k}-\mu)]$，$k=0,1,2,\cdots$，则自相关系数定义如下：

$$\gamma_k=\frac{\mathrm{Cov}(x_i,x_{i-k})}{\sqrt{\mathrm{Var}(x_i)}\sqrt{\mathrm{Var}(x_{i-k})}} \tag{5-29}$$

对于一个平稳的过程有：$\mathrm{Var}(x_i)=\mathrm{Var}(x_{i-k})=\sigma^2$，则公式（5-29）可以简化如下：

$$\gamma_k=\frac{Cov(x_i,x_{i-k})}{\sigma^2}=\frac{\rho_k}{\sigma^2}=\frac{\rho_k}{\rho_0} \tag{5-30}$$

当 $k=0$ 时，$\gamma_0=1$。

自相关系数表明了不同时期的数据之间相关程度，其取值范围在–1 到 1 之间，绝对值越接近于 1，则表明时间序列的自相关程度越高。

对于时间序列 y_t，在给定的 $y_{t-1}, y_{t-2}, \cdots, y_{t-k+1}$ 条件下，y_t 与 y_{t-k} 之间条件相关的相关程度用偏自相关系数 Φ_{kk} 来度量，且 $-1 \leqslant \Phi_{kk} \leqslant 1$，则有：

$$\Phi_{kk} = \begin{cases} r_1 & k = 1 \\ \dfrac{r_k - \sum_{j=1}^{k-1} \Phi_{k-1,j} \cdot r_{k-j}}{1 - \sum_{j=1}^{k-1} \Phi_{k-1,j} \cdot r_j} & k = 2,3,\cdots \end{cases} \tag{5-31}$$

式中，r_k 为滞后 k 期的自相关系数，$\Phi_{k,j} = \Phi_{k-1,i} - \Phi_{kk} \cdot \Phi_{k-1,k-j}, j = 1,2,\cdots,k-1$。

对于 ARMA(p, q)模型，其选择过程需要检验样本自相关函数和样本偏自相关函数的截尾性，具体过程一般如下：

（1）根据时间序列的散点图、自相关函数和偏自相关函数图，以 ADF单位根检验其方差、趋势及其季节性变化规律，对序列的平稳性进行识别。一般来讲，经济运行的时间序列都不是平稳序列；

（2）对非平稳序列进行平稳化处理。如果数据序列是非平稳的，并存在一定的增长或下降趋势，则需要对数据进行差分处理。如果数据存在异方差，则对数据进行处理，直到处理后数据的自相关函数值和偏相关函数值无明显地等于零；

（3）根据时间序列模型的识别规则建立相应的模型。若平稳序列的偏相关函数是截尾的，而自相关函数是拖尾的，则可断定该序列适合于AR 模型；若平稳序列的偏相关函数是拖尾的，而自相关函数是截尾的，则可断定该序列适合于 MA 模型；若平稳序列的偏相关函数和自相关函数都是拖尾的，则该序列适合于 ARMA 模型;

（4）确定 p，q 值大小，对参数进行估计。对 ARMA 模型输入不同 p，q 值，然后比较各个模型拟合度，通常采用赤池信息量准则（AIC）[14]和贝叶斯信息准则（BIC）[15]来评估模型拟合的优良，当 AIC 或 BIC 值越小，则模型拟合越优良。通过这些准则，确定最佳的 p，q 值，得到相应的参数估计；

（5）进行假设检验，诊断残差序列是否为白噪声。

5.4.2 预测算法

1．微博时间序列

下面以预测 30 天内微博的转发次数和覆盖范围为例。在 30 天内不同时刻的微博转发次数或覆盖范围可以构成相应的时间序列，其时间序列定义如下：

假设 t_n 为微博发布后的第 n 个时间间隔，x_{t_n} 是在 t_n 时间内微博转发次数，则微博转发次数时间序列定义如下：

$$X = \{x_{t_1}, x_{t_2}, \cdots, x_{t_n}\} \tag{5-32}$$

假设 t_n 为微博发布后的第 n 个时间间隔，c_{t_n} 是在 t_n 时间内新增加可观察到该微博的用户数目，则微博覆盖范围时间序列定义如下：

$$C = \{c_{t_1}, c_{t_2}, \cdots, c_{t_n}\} \tag{5-33}$$

对于微博转发次数时间序列，可以直接通过微博转发记录来构造。微博转发记录中包含有原始微博的 ID 号和发布时间以及微博转发的 ID 号和转发时间。因此，根据原始微博的 ID 号，可以收集到所有转发该微博的记录，然后从转发记录中提取转发时间来构建时间序列。

对于微博覆盖范围时间序列，无法直接通过微博转发记录来构造，需要与微博网络图结合起来。例如，可以从一条微博的所有转发记录中获取到不同时间段的用户数量及相应的用户 ID 号，因此某一时间段观察到该微博的新增加用户数目 c_{t_n} 是该时间段内所有转发用户的粉丝量的总和，而每个用户的粉丝量可以通过微博网络相应用户的入度数获取到。

2．算法实现

对于 M1 以及 M2，可以利用它们的时间序列曲线进行预测，首先利用开始时间段的微博转发次数或覆盖范围时间序列，对后续时间段（30 天内）的微博转发次数或覆盖范围时间序列进行预测，然后对所有时间片的值求和，最终得到预测值。因此，关键的问题是如何能准确地预测出后续时间段的时间序列曲线。

在图 5-11 中，t_1 时刻是前 100 条微博转发的截止时间点，实线表示开始时间段的微博转发时间序列曲线，虚线表示需要预测的微博转发时间序列曲线。

这里采用一种称为 RE_ARMA 的预测算法来预测微博转发时间序列，该算法基于 ARMA 模型思想，即时间序列中各个时间片并非完全独立，某一时刻的值与过去一个或多个时刻的值存在一定的关联性，因此可以利用过去一个或多个时刻的值来预测当前时刻的值。由于微博转发时间序列的时间片也不是完全独立的，例如某一时刻存在上升或下降的趋势，下一时刻的值会响应这一趋势，这说明两个时刻的值之间存在着关联性。

RE_ARMA 算法没有直接使用 ARMA 模型，因为开始时间段的微博转发时间序列太短而预测的时间序列太长，ARMA 模型很难准确地预测全部的时间序列。例如，大部分微博前 100 条转发时间都在前一个小时内，如果使用前一个小时的时间序列来训练 ARMA 模型，则会出现较大的误差。

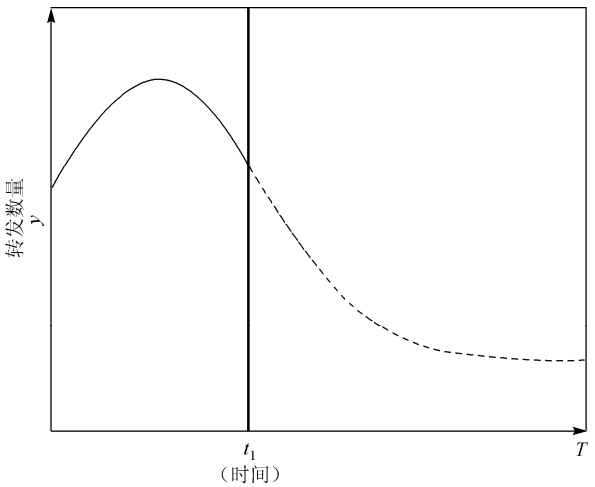


图 5-11　微博转发时间序列曲线

RE_ARMA 算法是从已有的历史数据中寻找与预测微博最相似的微博，也就是寻找与预测微博最相近的峰值和衰减指数，这是基于不同微博的转发时间序列曲线具有相似性。例如，微博转发时间序列一般都是在短时间内迅速地达到峰值，然后随着时间推移呈指数衰减直至最终消失，而这类曲线不同点在于峰值以及衰减指数大小。

RE_ARMA 算法主要有如下三个步骤：

（1）在历史微博数据中寻找与预测微博时间序列最相似的时间序列；

（2）利用 ARMA 模型训练该时间序列，学习相应的参数，再利用学习后的模型来预测微博转发；

（3）对 30 天内每个时间段的预测值进行求和，最终得到总的微博转发次数。

算法的第一步是最关键的，即如何高效地寻找最相似的微博转发时间序列。算法采用了两个方法：一是寻找相似微博的过程，采用欧拉距离来刻画微博相似性，如果两个微博的时间序列欧拉距离值越小，则这两个微博就越相似；二是从历史数据中寻找最相似微博的过程，每次都要计算欧拉距离，这一过程需要花费大量的计算时间。为了减少计算时间，可以采用贪婪式搜索过程，即每次从最相似的用户开始搜索，而判别用户相似性的依据是两个用户的粉丝数。算法优先寻找粉丝量相近用户的微博，计算欧拉距离，如果欧拉距离较大，则继续寻找下一个用户的微博，直到欧拉距离小于一定的值，则停止搜索。

算法的第二步和第三步比较简单，对已经寻找到的微博转发时间序列建立相应 ARMA(p, q)模型，包括时间序列的平稳性分析、模型选择以及参数 p 和 q 设定等，然后使用该模型对微博进行预测，最终对 30 天内的微博转发时间序列的值相加求和，得到最终预测值。

RE_ARMA 算法既可用于预测微博转发次数，又可用于预测微博覆盖范围，算法的具体操作过程如下。

算法 5-3　微博转发次数预测算法（RE_ARMA）

输入：训练数据集 $v=\{D_1,D_2,\cdots,D_N\}$

预测微博的已知部分时间序列 $X=\{x_1,x_2,\cdots,x_T\}$

输出：预测 30 天微博转发数 $\sum_{t=1}^{M}x_t$，其中 M 是第 30 天的时间点。

（1）贪婪式搜索时间序列 $D\in v$，使欧拉距离 $\sum_{t=1}^{T}(d_t-x_t)^2$ 最小；

（2）ARMA(p, q)模型拟合 $D=\{d_1,d_2,\cdots,d_N\}$；

（3）在 X 中采用拟合的 ARMA（p, q）模型预测 $z_t(t=T+1,\cdots,N)$；

（4）输出 30 天的转发总数 $\sum_{t=1}^{M}x_t$。

5.4.3　算法验证

下面通过实验数据对微博转发特性预测算法性能进行测试和验证。

1．实验数据集

实验数据来自新浪微博，数据集中分为训练数据集和测试数据集两个部分。

1）训练数据集

训练数据集主要包括两个部分数据：微博转发记录和微博关注网。

对于微博转发记录数据，共采集了 5636858 个用户发布的 46584914 条原始微博和相应的 190920026 条转发记录，平均每个用户和每条微博转发次数分别为 34 次和 4 次。微博转发次数分布如图 5-12 所示，从图 5-12 可以看出，微博转发次数分布符合幂律分布，大部分原始微博转发次数较少，而小部分原始微博转发次数较多。曲线斜率约为 2，表明转发次数少的微博比例较大，且微博数目随着转发次数呈指数剧烈地下降，例如，转发次数不到 10 次的微博高达 95%，而大于 10 次的微博不到 5%。同时它还有较长的拖尾，说明存在转发次数较大的微博，而微博最大的转发次数达到了 30389 次。

对于微博关注网数据，共收集了 48234064 个用户节点和 198340762 条关注边。图 5-13 和图 5-14 分别给出了关注网的出度和入度分布。由图 5-13 可见，出度曲线近似为一条直线，斜率约为 2.5，符合幂律分布，说明大部分用户的关注边较少，而小部分用户的关注边较多，符合真实情况。由于大部分用户投入的时间和精力有限，因此关注人数是有限的，而一小部分用户花费大量的时间，长期活跃在微博中，关注人数众多。

在图 5-14 中的入度曲线则稍有不同，虽然大体上也是大部分用户入度数较小，而小部分用户入度数较大，但是并不符合幂律分布下降。当度数大于 100 时，曲线变得平

缓，且在度数约为 1000 处又开始急剧地下降，说明入度数为 100 至 1000 之间的用户数目分布较为均匀。另外，在度数约为 8000 处，曲线有一个峰值，这说明此处的用户数目较多。

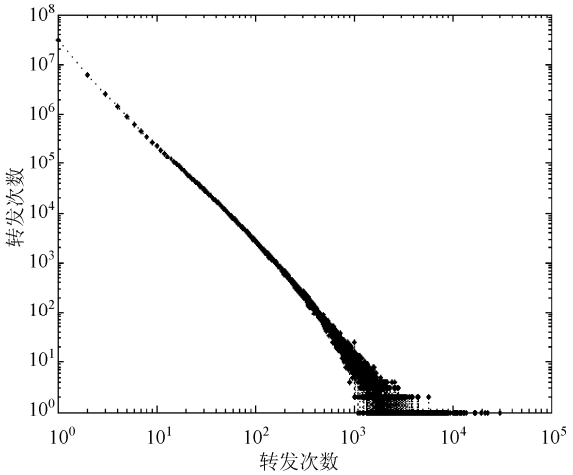


图 5-12　微博转发次数分布

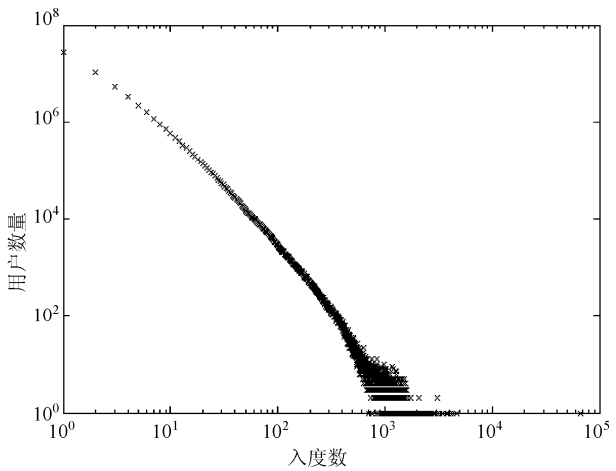


图 5-13　微博关注网的出度分布

转发记录中用户的 ID 号与关注网的用户 ID 号是一一对应的，因此，对于微博的覆盖人群数，可以依据微博转发记录和用户关注网的 ID 号统计出来。例如，微博转发用户可以通过 ID 号在关注网找到其粉丝，然后统计所有转发用户的粉丝之和。由于微博转发记录中用户 ID 号与关注网中用户 ID 号并不是完全重合，也就是说，有可能微博转发记录中的用户并不存在于关注网中。为了弥补这个问题，用户的粉丝量可以用设置的默认值来替代，例如，当一用户的 ID 号不在关注网中，用均值粉丝量来替代。

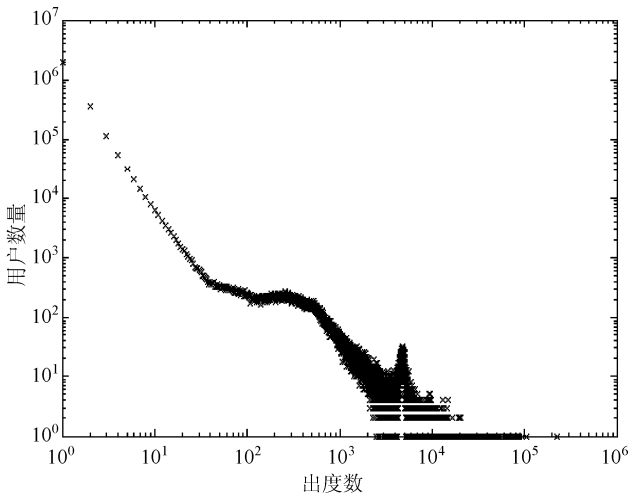


图 5-14　微博关注网的入度分布

2）测试数据集

测试数据集的微博主题内容分布如表 5-10 所示，其中包括 6 个不同类型主题的热点事件，这说明不同微博内容是不一样的，涵盖不同主题内容。

表 5-10　测试数据集的微博主题分布

主题名	消息数	主题名	消息数
乔布斯逝世	4	李娜赢得法网公开赛冠军	5
福州爆炸案	5	小米手机发布	5
日本大地震	6	药家鑫谋杀事件	8

经过统计发现，发布微博的用户分别来自于 27 个不同的用户，其中用户的粉丝量如图 5-15 所示，不同用户的粉丝量差异较大，最大相差达到 1000 倍以上，这说明测试数据集中的用户类型并不完全相同，涵盖不同类型的用户。另外，还对测试数据集的前 100 条微博转发所需时间进行了统计，如图 5-16 所示，大部分微博前 100 条转发时间都比较短，不到一个小时，这符合微博时效性强的特点，同时也存在前 100 条转发时间较长的微博，最长达到了 10 天。可见，测试数据集涵盖了不同长度的转发时间。因此，33 条测试数据包括了不同类型主题、不同类型用户以及不同转发时间的微博，这说明数据具有代表性，可以用于对预测模型的科学评估。

2．算法性能对比

SPSS 是一个比较流行的统计分析和绘图软件，可用来统计数据分析结果。模型拟合优良性评估通常采用赤池信息量准则（AIC）和贝叶斯信息准则（BIC），AIC 或 BIC 值

越小，模型拟合则越优良。SPSS 软件提供了标准的 BIC，因此将 BIC 作为模型参数选择的主要依据。另外，算法还要综合考虑其他的准则，如 R-Square、均方根误差等。

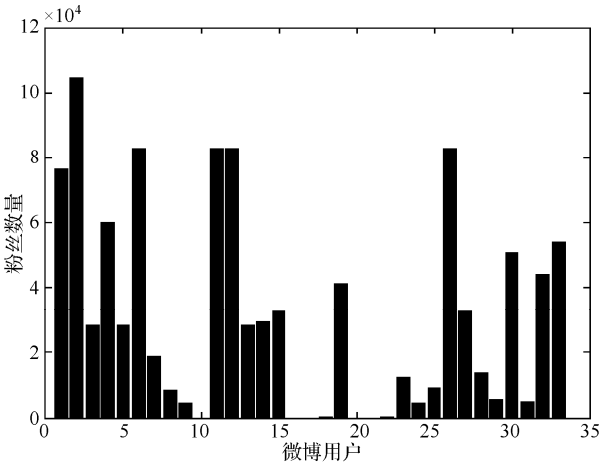


图 5-15　测试数据集的微博粉丝分布

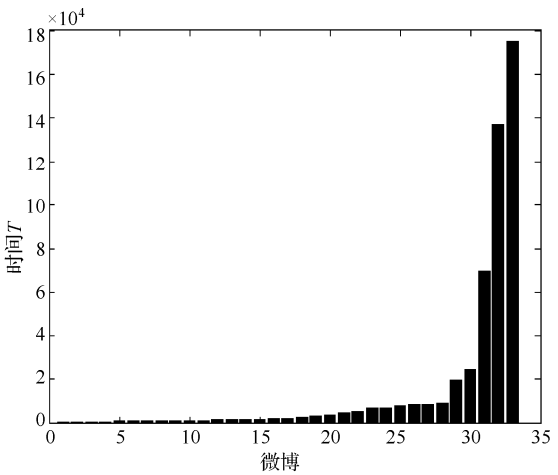


图 5-16　测试数据集的微博转发时间分布

算法首先使用训练数据集来学习参数和训练模型，然后对 33 个测试数据进行预测，测试的 M1 值和 M2 值分别如图 5-17 和图 5-18 所示。在 M1 值预测中，每条微博的平均转发次数为 487 次，其中转发次数最多的微博是一条关于乔布斯逝世的微博，达到 3862 条，而转发次数最少的微博则是一条关于小米手机发布的微博，只有 107 条。在 M2 值预测中，每条微博的平均覆盖范围（人数）为 147831 个用户，其中覆盖人数最多的微博是一

条关于药家鑫谋杀事件的微博，总共 1428387 个用户，而覆盖人数最少的微博也是关于小米手机发布的微博，只有 4200 个用户。由此可见，不同微博的 M2 值比 M1 值的差异性更大，因此 M2 值预测难度也更大。另外，微博的 M1 值与 M2 值大小并不完全一致。例如，在第 26 条测试数据的预测中，M1 值较小，但是 M2 值却很大，这也从侧面说明了微博的转发数与覆盖范围意义是不同的。

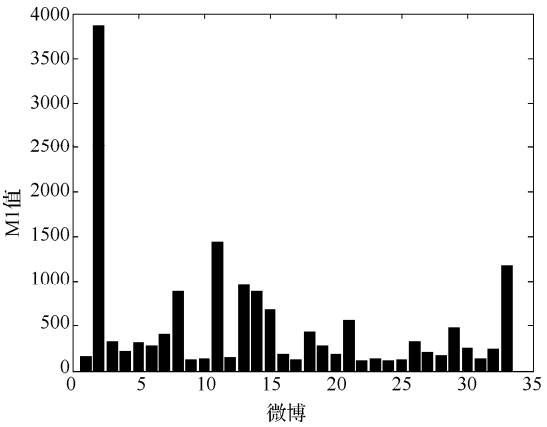


图 5-17　预测微博转发次数分布

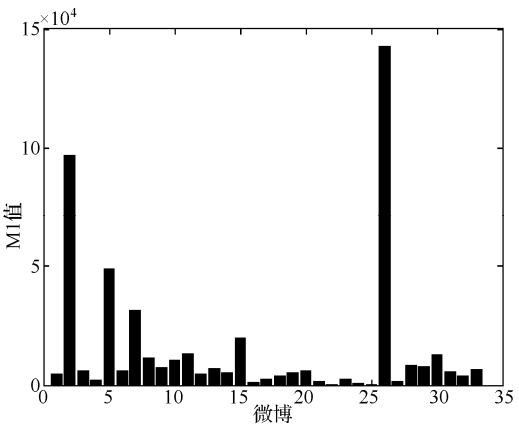


图 5-18　预测微博覆盖范围分布

这里采用计算预测值与实际值的差值大小作为评价分的方法来评价预测结果的优劣，评价分计算公式如下：

$$s_i = \frac{|p_i - r_i|}{r_i} \tag{5-34}$$

式中，p_i 表示微博预测转发次数或覆盖范围，r_i 表示微博实际转发次数或覆盖范围。由公式（5-34）可知，微博的评价分越低，则该条微博预测结果就越准确。

对所有微博而言，期望的评价分计算公式如下：

$$S=\frac{1}{n}\sum_{i=1}^{n}s_i \tag{5-35}$$

当 S 值越小，则微博的预测值将与实际值越接近，预测结果就越理想。

表 5-11 给出了几种预测算法的预测结果对比，由表5-11可见，RE_ARMA 和 RE_CM 算法对微博转发次数和覆盖范围的预测结果比较接近，两者的平均评价分相差不大，表明这两种算法的预测结果比较准确，而其他预测算法的预测结果较差。由于 RE_CM 算法在预测过程采用了多个机器学习方法[16]，预测过程比较复杂，算法复杂度较高，而 RE_ARMA 算法则比较简单，算法复杂度低，也容易理解。

表 5-11　几种预测算法的预测结果对比

算法	#Retweet	#Possible View	平均评价分
RE_CM	20.7	21.6	21.15
RE_ARMA	27.5	25.2	26.35
RE_CF	67.2	73.23	70.2
RE_LR	297.1	290.3	243.7
RE_RW	332.5	1221.1	776.8

综上所述，微博转发时间序列曲线具有相似性，这种时间序列相似性可以用于预测微博转发次数和覆盖范围，预测过程可以看作是从过去的时间序列中寻找最相似微博的过程。RE_ARMA 算法是一种基于 ARMA 模型的转发预测算法，能够准确地预测微博转发次数和覆盖范围，并且简单高效。

5.5　微博转发峰值分析

微博具有很强的时效性，不同时刻的微博关注度是不同的，关注度的时间序列反映了微博受欢迎程度的变化，在某一时刻的微博关注度达到峰值则表明了此时的微博最受用户欢迎和关注。因此，时间序列峰值是微博转发时间序列的最重要特征，对于预测微博转发行为和特性具有重要的意义。

5.5.1　时间序列概念

微博转发时间序列是指微博转发次数随着时间变化的曲线，下面给出相关时间序列的概念和定义。

给定 t_n 为微博发布后的第 n 个时间间隔，x_{t_n} 是在 t_n 时段内微博转发数，则微博转发时间序列定义为：

$$X=\{x_{t_1},x_{t_2},\cdots,x_{t_n}\} \tag{5-36}$$

假定阈值 n_p 是微博转发时间序列峰值的临界点，则在峰值 X_p 中有：

$$x_{t_k}\begin{cases}\geqslant n_p & t_k\in[t_i,t_j]\\ <n_p & t_k\in t_{i-1},t_{j+1}\end{cases} \tag{5-37}$$

式中，x_{t_k} 表示在峰值 X_p 内的转发次数，t_i、t_j 分别为峰值 X_p 的开始时间和结束时间。峰值 X_p 的时段长度 T_p 定义为：

$$T_p=t_j-t_{i-1} \tag{5-38}$$

假定 T_{p_i} 是微博转发时间序列的第 i 个峰值的时段长度，则总峰值时间定义为：

$$T_{\text{pall}}=\sum T_{p_i} \tag{5-39}$$

总峰值时间是微博转发时间序列所有峰值时段之和。

给定 $t_0=0$ 是原始微博的发布时间，t_e 是微博转发时间序列最后一个峰值的结束时间，则微博存活时间定义为：

$$T_d=t_e-t_0=t_e \tag{5-40}$$

微博存活时间是微博整个生存周期，其中包括了峰值时段和非峰值时段。

图 5-19 为微博转发时间序列示例图，该微博转发时间序列存在 3 个峰值，其中每个峰值幅度以及时段长度都不一样。总峰值时间是 3 个峰值时段长度之和，而存活时间是指微博发布时间开始到最后一个峰值结束这一时间段。

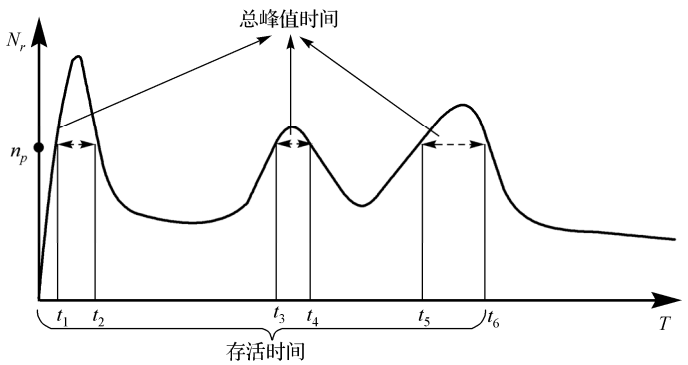


图 5-19　微博转发时间序列示例

由此可见，总峰值时间与存活时间并不一样。总峰值时间主要用来描述微博欢迎度最高的时间段长短，而存活时间则描述微博的存活时间长短。

给定 $A \to B \to \cdots J \to K$ 是一条微博转发路径，N_{re} 是该转发路径上用户转发次数，则微博转发路径长度定义为：

$$P_l = N_{\mathrm{re}} \tag{5-41}$$

微博转发路径长度反映了微博传播的深度，是从网络结构上来研究微博转发的性质。

5.5.2 峰值检测算法

定理 5-3：（切比雪夫不等式）给定一个随机变量 X，期望值为 μ，标准方差为 σ。则对于任何实数 $\lambda > 0$ 有：

$$P_r(|X - \mu| > \lambda\sigma) \leqslant 1/\lambda^2 \tag{5-42}$$

切比雪夫不等式表明了在任何样本数据或者概率分布中，几乎所有的值都“接近”于均值，更精确地表述是不超过 $1/\lambda^2$ 的数据偏离均值 μ 的 $\lambda\sigma$ 距离。由此可见，大部分数据都处在均值 μ 的上下 $\lambda\sigma$ 范围内分布，只有小数离群点偏离均值较远。因此，由切比雪夫不等式可知，在微博转发时间序列中每个时间片的值都在均值线上下一定范围分布，而一个显然的事实是峰值为偏离均值较远的“离群点”。

这里的主要问题是如何在微博转发时间序列中寻找到这些“离群点”，并如何确定峰值阈值大小，而阈值大小将决定时间序列中峰值的个数、峰值时段长度以及总峰值时间等。阈值设置太小，将导致过多的伪峰值；阈值设置过大，将导致一些峰值检测不到。文献[17]通过切比雪夫不等式的上界来确定峰值阈值 $N_p = \mu + \lambda\sigma$，即任何转发数大于 N_p 的时间片都被看作是峰值。该方法需要确定 λ 值，λ 值越大，比重 $1/\lambda^2$ 的数据越少，反之亦然。可见，λ 值决定“离群点”所占比例。在统计学中，小于 5%的事件通常被认为是小概率事件，因此该方法定义了比例小于 5%的数据是“离群点”，由此可得到 λ 值。该方法存在的问题是所求解的“离群点”不但包括上界的峰值点，也包括下界的谷底点，并不完全是峰值点。这个问题可以采用切比雪夫不等式的变形不等式（即 Cantelli 不等式）来解决，使之只识别单边的上界“离群点”。

定理 5-4：（Cantelli 不等式）给定一个随机变量 X，期望值为 μ，标准方差为 σ。则对于任何实数 λ 有：

$$P_r(X - u \geqslant \lambda\sigma) \begin{cases} \leqslant 1/(1+\lambda^2) & \lambda > 0 \\ \geqslant 1 - 1/(1+\lambda^2) & \lambda < 0 \end{cases} \tag{5-43}$$

Cantelli 不等式是单边的切比雪夫不等式变形，用于评估样本数据大于或者小于均值

的概率。由于时间序列的峰值高于期望值，则 $\lambda > 0$，由公式 5-43 可推导出：

$$P_r(X - u \geqslant \lambda\sigma) \leqslant 1/(1+\lambda^2) \tag{5-44}$$

由公式（5-44）可知，对于任何 $X > u + \lambda\rho$ 的数据，所占比例不超过 $1/(1+\lambda^2)$。定义的“离群点”是比例小于 5%的数据，由此得到 $\lambda \approx 4.3589$。算法 5-4 给出了峰值检测算法的具体步骤如下。

算法 5-4　微博转发时间序列波峰检测算法（RT_PEAK）

输入：时间序列 $X = (x_1, x_2, \cdots, x_n)$

概率 $P_r(X)$

输出：峰值数据集 o

（1）　$o = 0$；

（2）　$\bar{X} = \sum_{i=1}^{N} x_i \Big/ N$；

（3）　$\sigma^2 = \sum_{i=1}^{N} (x_i - \bar{X})^2 \Big/ N$；

（4）　$\lambda = (1 - P_r(X)/P_r(X))^{-2}$；

（5）　int peak_begin，　peak_length = 0;

（6）　For(int i = 1;i<N;i++)do

（7）　　If($x_i > \mu + \lambda\sigma$)then

（8）　　　peak_length++;

（9）　　　peak_begin = i- peak_length;

（10）　　Else

（11）　　　If（peak_length>0）

（12）　　　　$o = o \cup$ (peak_begin, peak_length);

（13）　　　　peak_length = 0;

（14）　　　End if

（15）　End for

5.5.3　峰值特性分析

下面通过实验数据对微博转发峰值特性进行分析。

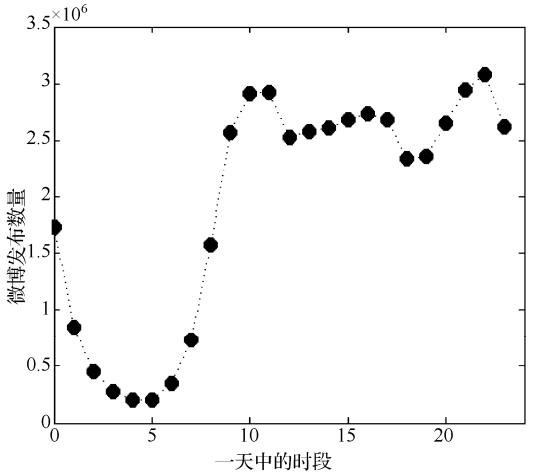
1．实验数据集

实验数据来源于新浪微博，数据集中包含了微博转发记录，微博转发记录包括了 5636858 个用户发布的 46584914 条微博以及相应的 190920026 条转发记录。

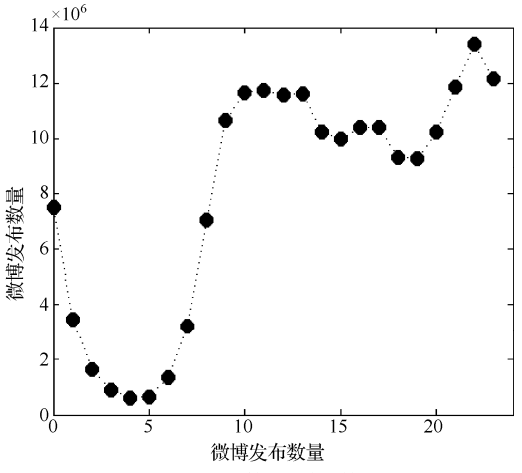
2．微博时间特性

在微博发布或转发过程中存在着周期性。例如，在一天中的不同时间段用户发布或转发微博数目是不同的。以 24 小时作为周期，图 5-20 给出了不同时段微博发布和转发次数

的分布，微博发布和转发次数的曲线变化大体上相同，且与人们一天的作息规律相一致。例如，从早上 8 点开始，用户开始活动，发布和转发微博次数逐渐增加，并一直持续在较高的区间，直到凌晨零点后才下降。在这一区间出现了两个峰值，第一个峰值出现在上午 10 点左右，持续时间达 2 个小时；第二个峰值出现在晚上 9 点左右，持续时间达 3 个小时，也是一天中最高的峰值，表明在这两个时间段微博用户最多，尤其是晚上时间段。在中午 12 点和下午 6 点左右，曲线都会出现一个小波谷，表明在这两个时间段用户在用餐或休息。从凌晨 0 点开始至早上 8 点前，微博发布和转发数是一天中最少的，其中最低谷出现在凌晨 5 点，不到最大峰值的 1/20。



（a）用户发布微博次数



（b）用户转发微博次数

图 5-20　24 小时微博发布和转发次数分布图

另一个问题是微博的时效性，微博的时效性可以看作是用户对微博关注度随着时间变化过程，微博关注度可以用微博转发次数来表征，例如某一时刻微博转发次数越多，表明此时微博的关注度也越高。因此，微博的时效性主要是研究微博转发次数的时间序列变化。

图 5-21 给出了微博转发时间间隔分布，符合幂律分布，大部分微博的转发时间间隔较短，具有很强的时效性，只有小部分微博的转发时间间隔较长。例如，81.8%的微博在一天内被转发，而转发时间超过两天的微博数目不足 12%，并且随着时间推移，其微博数量呈指数下降。这说明微博具有很强的时效性。

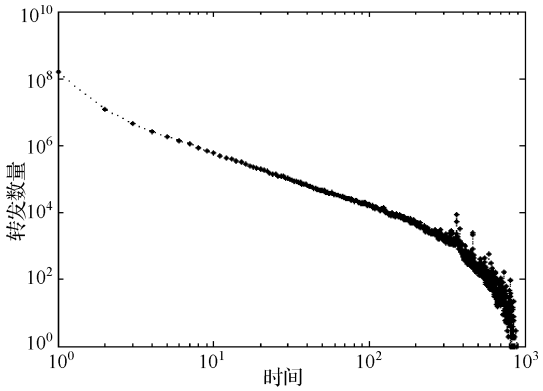


图 5-21　微博转发时间间隔分布

3．微博转发峰值检测

为了更好地分析时间序列峰值，从数据集中检测出转发次数超过 100 的微博峰值分布，总共得到 207259 条微博及相应的 48302776 次转发记录，平均每条微博转发次数约为 240 次，然后将这些微博转化为相应的转发时间序列，每一个时间刻度为 6 小时。

微博峰值检测结果如表 5-12 所示，其中没有峰值的微博有 26620 条，大约占 12.84%，说明微博网络存在大量不活跃的微博；含有一个峰值的微博有 142406 条，大约占 68.71%，属于正常的微博；含有多个峰值的微博有 38233 条，大约占 18.45%，有些微博的峰值甚至达到 12 个，说明微博容易受到其他因素影响而导致多次被高度关注。

表 5-12　不同峰值的微博数目及比例

峰值个数	微博数目	比例
0	26620	12.84%
1	142406	68.71%
≥2	38233	18.45%

表 5-13 给出了不同峰值的微博的平均转发路径长度，其中峰值为 0 和 1 的微博的平均路径长度比较接近，说明这两类微博的传播深度相类似。多个峰值的微博的平均路径长度达到了 2.32，说明多个峰值的微博的转发路径较长，传播深度较深。

表 5-13 不同峰值的微博平均转发路径长度

峰值个数	平均转发路径长度
0	2.03
1	2.09
≥2	2.32

4．微博转发峰值分析

微博转发峰值分析主要针对不同类型热门主题和用户的微博，分析这些微博的各个特征，包括总峰值时间、存活时间、转发路径长度以及峰值时段路径长度等。

1）微博主题类型分析

微博中包括了 44 个不同的热门主题，这里提取了转发次数前 4 位的主题微博进行分析，它们是房价问题事件、河北大学校园飙车致死案事件、李阳家暴事件以及小米手机发布事件。另外，对于不含任何热门主题的微博，也看作是一种类型的微博。

表 5-14 给出了不同类型主题微博的特征比较，它们有两个特点，一是含有热门主题和不含有热门主题微博特征存在较大的差异，不含热门主题微博的转发路径长度（P_i）、总峰值时间（T_p）及存活时间（T_d）等都较短，转发路径长度只有 1.90，大约为含有热门主题微博的转发路径长度的 60%～70%，这说明不含热门主题的微博被关注时间和存活时间都较短，传播范围有限。二是不同类型热门主题的微博特征也不同，例如，河北大学飙车致死事件的转发路径和存活时间较长，而小米手机发布事件的转发路径和存活时间则较短，这说明河北大学飙车致死事件的社会影响大，受到用户的高度关注，具有较长的存活时间。

表 5-14 不同类型主题微博的特征比较

主题类型	Average(P_I)	Average(T_p)	Average(T_d)(days)
不含任何热门主题	1.90	2.95	15.85
房价问题	2.66	3.05	22.75
小米手机发布	2.62	3.39	16.32
李阳家暴事件	2.89	3.71	21.20
河北大学飙车致死事件	2.97	3.64	28.15

图 5-22 给出了不同主题微博的转发路径分布图，随着转发路径长度增加，不同主题微博所占比率都在下降，但是下降幅度有所不同。在不含热门主题微博中，转发路径较短

的微博比率较大，其中路径长度小于 3 的微博比率接近 84%；而路径长度大于 5 的微博比率不到 2%，呈现急剧下降趋势。在含有热门主题微博中，下降趋势较为平缓，转发路径较长的微博比率较大，转发路径大于 5 的微博比率达到 12%以上。

图 5-23 给出了不同主题微博的总峰值持续时间分布，所有微博的总峰值时间前 4 个所占比率最大且不含热门主题的微博比率最大，占到总数的 98%。含有热门主题的微博下降趋势较为平缓，并且还存在小部分总峰值时间较长的微博，最大达到 15 个单位时间，这说明了此类型热门主题处于高度关注的时间较长。

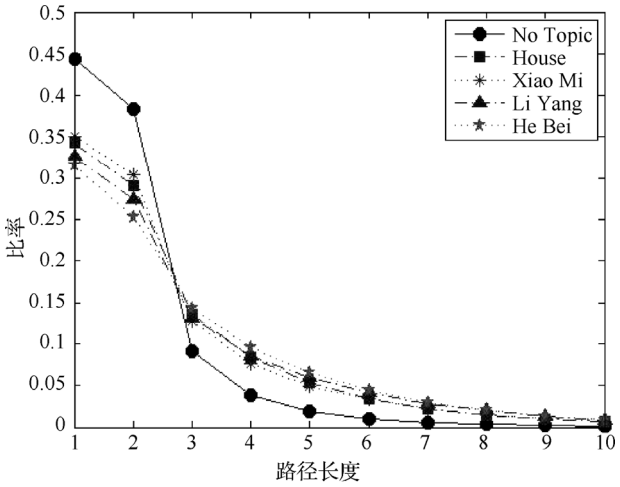


图 5-22　不同主题微博的转发路径分布

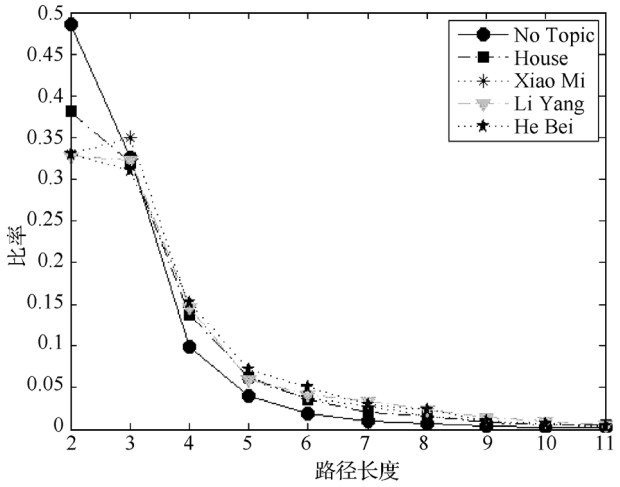


图 5-23　不同主题微博的总峰值时间分布

表 5-15 给出了不同主题类型微博的存活时间，所有微博的存活时间都呈现下降趋势，但不同主题的微博下降幅度有所不同。在不含热门主题的微博中，存活时间较短，存活时间在 3 天内的微博比率达到 75%，3 天后迅速下降，之后下降幅度又变缓，存在较长的拖尾，这说明存在一部分存活时间较长的不含热门主题微博。在含有热门主题的微博中，存活时间较长，尤其是河北大学飙车致死事件，存活时间低于 3 天的微博比率只有 60%左右，与不含热门主题的微博相差 15%，而存活时间超过 100 天以上的微博比率达到 9.45%，几乎是不含热门主题的微博的 2 倍，这说明此类主题的微博存活时间长，具有很强的生命力。另外，房价问题、李阳家暴事件等热门主题的微博同样具有较强的生命力，而小米手机发布事件的微博存活时间与不含热门主题的微博相似，说明此类微博的生命力一般。

表 5-15 不同类型主题微博的存活时间

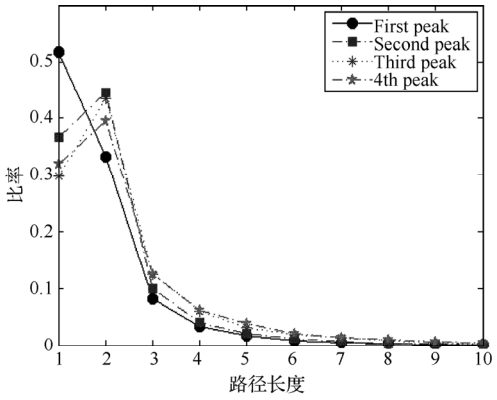
	不含主题		房价问题		小米手机发布		李阳家暴事件		河北大学飙车致死事件	
天数	数目	比率	数目	比率	数目	比率	数目	比率	数目	比率
1	11437	47.55	2369	38.31	2109	44.83	1110	40.20	524	34.13
2	5477	22.77	1362	22.02	1067	22.68	628	22.74	329	21.43
3	1126	4.68	372	6.01	252	5.35	165	5.98	87	5.67
4～10	1995	8.29	684	11.06	451	9.59	275	9.96	180	11.73
10～100	2873	11.95	937	15.15	586	12.46	398	14.42	270	17.59
>100	1145	4.76	461	7.45	239	5.09	185	6.70	145	9.45
总共	24051	100	6185	100	4704	100	2761	100	1535	100

表 5-16 给出了不同峰值转发路径长度分布，它有两个显著的特征，一是在相同的主题微博中处于后面峰值的微博比处于前面峰值的转发路径长度要长，这说明随着时间的增加，微博转发路径在增长，深度在增加；二是在不同峰值时段含有热门主题微博的转发路径长度比不含热门主题的微博要长，说明含有热门主题的微博具有更大的影响力。

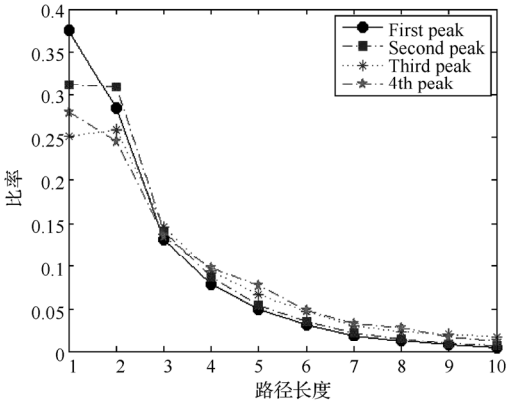
表 5-16 不同峰值转发路径长度分布

主题类型	第一个峰值	第二个峰值	第三个峰值	第四个峰值
不含任何主题	1.78	2.01	2.32	2.39
房价问题	2.51	2.71	3.53	3.30
小米手机发布	2.46	2.65	3.71	3.45
李阳家暴事件	2.69	2.92	4.09	3.80
河北大学飙车致死事件	2.76	3.04	4.04	3.64

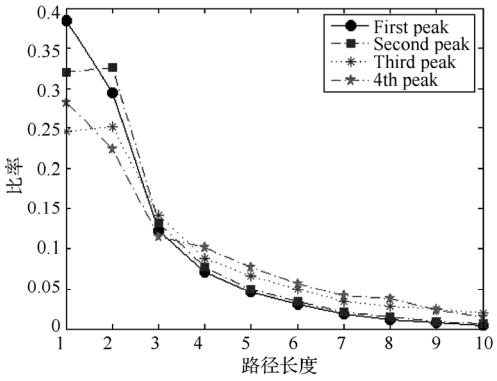
图 5-24 给出了 5 种不同类型微博在不同峰值时段转发路径长度比率分布。在不含热门主题的微博中，第一个峰值转发路径长度为 1 的微博比率最大，达到了 50%，其他峰值转发路径长度为 1 的微博比率下降幅度较大，这说明在不含有热门主题的微博中，第一个峰值的转发方式与其他峰值不同。例如，在第一个峰值中大多数用户可能直接转发原始微博，而在其他峰值中用户转发的微博可能来自于其他用户。在含有热门主题的微博中，各个峰值的转发路径比率曲线变化比较相似，这说明了不同峰值时段的转发方式比较相似。



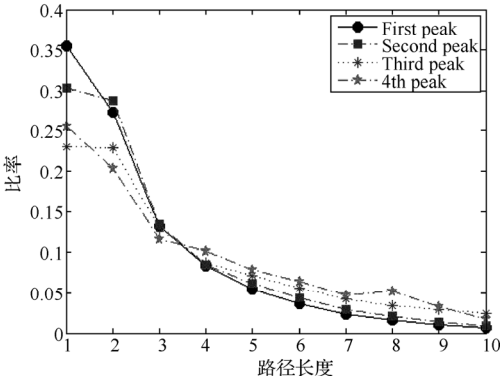
（a）不含热门主题微博



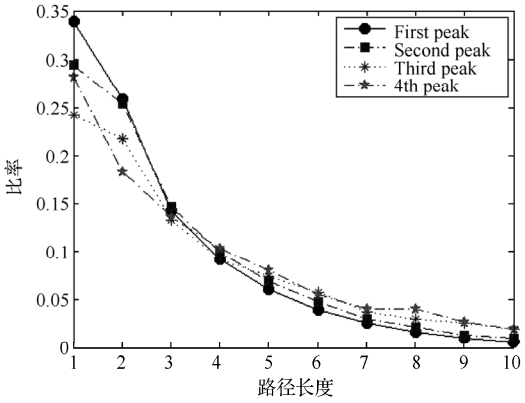
（b）房价问题事件



（c）李阳家暴事件



（d）小米手机发布事件



（e）河北大学飙车致死事件

图 5-24　不同类型主题微博在不同峰值转发路径比率分布

通过以上分析可知，不同类型的主题对微博转发时间序列多峰值有较大影响，具体表现在热门主题微博的转发路径长度、总峰值时间、存活时间以及峰值转发路径长度比不含热门主题的微博更长，这说明热门主题的微博影响力大、关注时间长以及生命周期长，并且不同热门主题的微博表现也不完全相同。

2）微博用户类型分析

用户类型包括超级用户、次超级用户和一般用户。不同类型用户的微博特征如表 5-17 所示，超级用户的微博转发路径长度、总峰值时间以及存活时间都较短，与一般用户微博相比，差距比较明显。超级用户的微博转发路径长度只有 1.86，大约是一般用户微博转发路径长度的 61%，而超级用户的存活时间大约是一般用户存活时间的 37%。次超级用户则处于超级用户与一般用户之间。

表 5-17　不同类型用户的微博特征比较

用户类型	Average(P_l)	Average(T_{pall})	Average(T_d)(days)
超级用户	1.86	2.73	10.63
次超级用户	2.17	3.05	18.34
一般用户	3.04	3.77	28.46

图 5-25 给出了不同类型用户微博转发路径长度分布，随着转发路径长度增加，所有微博比率都呈现下降趋势，不同类型用户的微博比率下降幅度有所不同，在超级用户和次超级用户中，路径长度短的微博比率较大，路径长度小于 2 的微博比率接近 80%，并且随着路径长度增长微博比率下降明显。在一般用户中，路径长度短的微博比率下降明显，路径长度小于 2 的微博比率下降幅度达到 25%，而路径长度长的微博比率有所上升，整个曲线较为平缓。

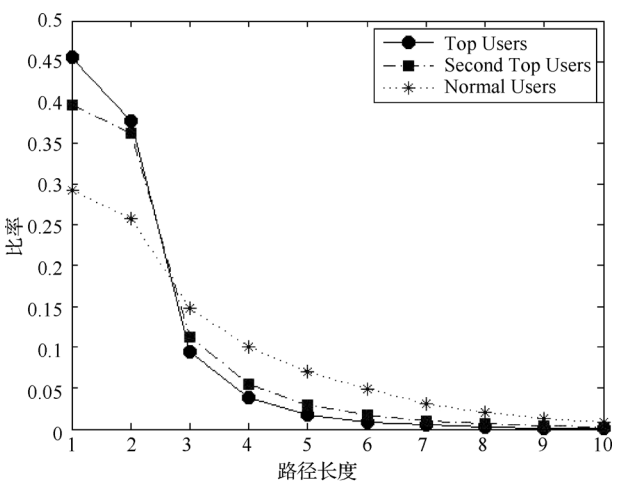


图 5-25　不同类型用户微博转发路径长度分布

图 5-26 给出了不同类型用户的微博总峰值时间分布，随着峰值持续时间增长，所有微博比率迅速下降，且总峰值时间中前 4 个所占比例最大。与图 5-25 相类似，超级用户的下降幅度比较剧烈，而一般用户的下降幅度较为平缓。

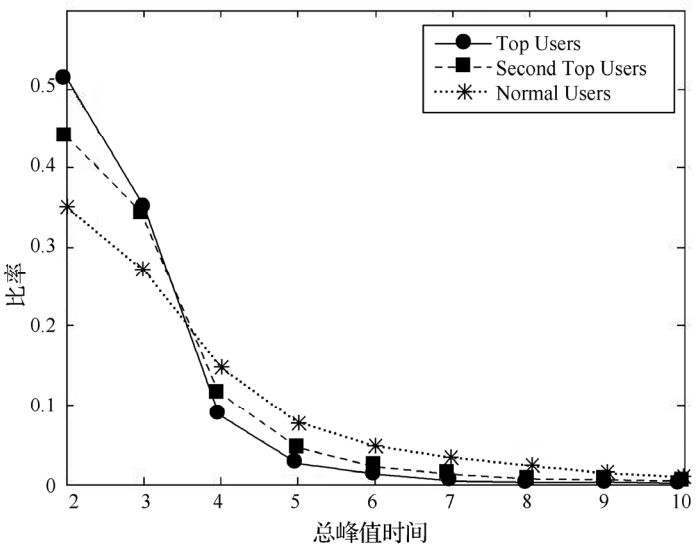


图 5-26　不同类型用户微博总峰值时间分布

表 5-18 给出了不同类型用户微博的存活时间分布及比率，随着存活时间的增长，所有微博比率都呈现下降趋势，但不同类型用户的微博比率下降幅度有所不同。在超级用户中，微博的存活时间都较短，存活不超过 1 天的微博比率达到 60.25%，大大超过了平均值；存活超过 3 天的微博比率较少，不足 16%。在一般用户中，情况则不同，存活时间短的微博比率大幅度减少，而存活时间较长的微博比率较高，例如存活不超过 1 天的微博比率下降到 23.21%，存活超过 3 天的微博比率达到 54%，存活 100 天以上的微博比率高达 8.74%。由此可见，超级用户的微博存活时间分布与一般用户完全不同，次超级用户则处于超级用户与一般用户之间。

表 5-19 给出了用户在不同峰值时段转发路径长度，它有两个特征，一是在相同类型用户中微博后一个峰值的转发路径长度比前一个峰值要长，尤其在超级用户中，这一特征更加明显，例如第 4 个峰值时的转发路径长度比第一个峰值长近 50%，这说明超级用户在不同峰值的转发方式是不同的；二是在不同峰值时段一般用户的微博转发路径长度都比超级用户要长，说明一般用户的微博传播深度更深一些。

表 5-18　不同类型用户微博存活时间分布及比率

	超级用户		次超级用户		一般用户	
天数	数目	比率	数目	比率	数目	比率
1	10055	60.25	4964	43.19	2334	23.21
2	3998	23.96	2633	22.91	1930	19.20
3	461	2.76	608	5.29	840	8.36
4～10	679	4.07	1040	9.05	1717	17.08
10～100	944	5.66	1602	13.94	2352	23.40
>100	551	3.30	646	5.62	879	8.74
总数	16688	100	11493	100	10052	1000

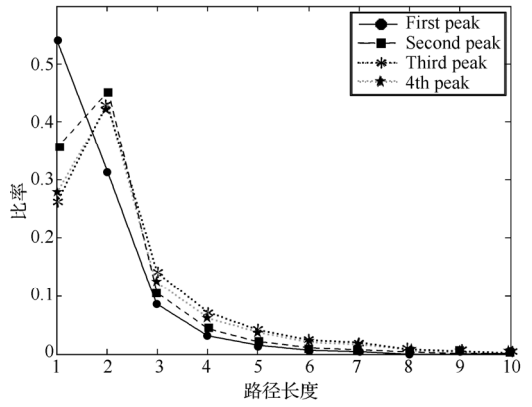
表 5-19　用户在不同峰值时段转发路径长度

	第一个峰值	第二个峰值	第三个峰值	第四个峰值
超级用户	1.71	2.01	2.48	2.53
次超级用户	2.02	2.23	2.95	2.86
一般用户	2.99	3.03	3.22	3.26

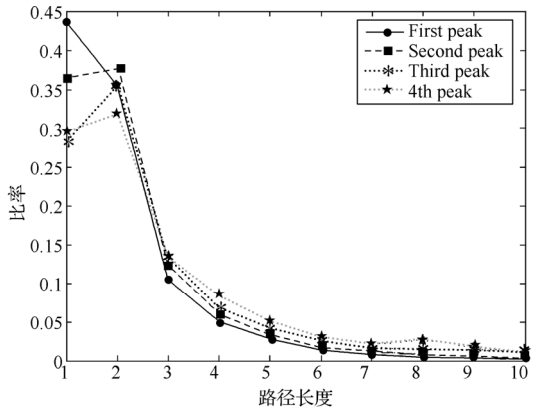
图 5-27 中给出了三种类型用户在不同峰值时段转发路径长度比率分布。在超级用户中，第一个峰值转发路径长度为 1 的微博比率很大，达到 50%，其他峰值转发路径长度为 1 的微博比率下降幅度较大，这再次说明了超级用户在第一个峰值的转发方式与其他峰值的转发方式不同。在一般用户中，不同峰值的转发路径比率变化比较相似，几乎重叠，这说明了用户的微博在不同峰值的转发方式基本相似。由此可见，不同类型用户的微博产生多峰值的原因是不同的。

以上数据分析表明，在不同类型的用户中，多个峰值的微博具有不同的特征，超级用户的微博转发数量较多，在短时间内快速地转发，其转发路径长度、总峰值时间以及存活时间都较短，因此超级用户的微博生命力偏短，可以看作是一种“爆炸式”传播方式。在一般用户中，虽然微博转发数量较少，转发速度较慢，但是转发路径长度、总峰值时间以及存活时间都较长，对路径长度大于 2 的用户影响力更大，因此可以看作是一种“蔓延式”的传播方式。而次超级用户兼有两者的特征，既具有超级用户的“爆炸式”传播方式，也具有一般用户的“蔓延式”传播方式。

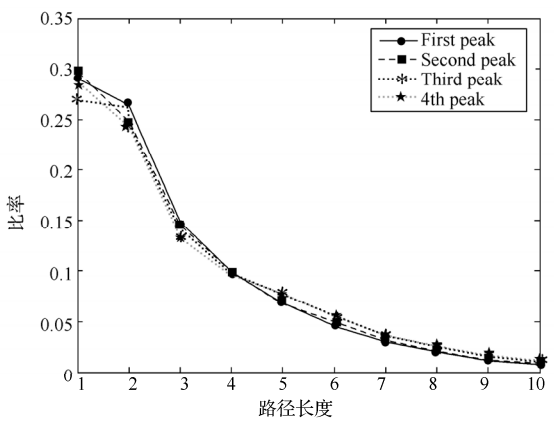
对于超级用户和一般用户不同的微博转发方式，主要是由微博网络特殊性所决定的，他们在网络中所担当的角色不一样。超级用户类似于媒体网络中的媒体，向用户传播信息主要依靠基于服从权威所形成的影响力，传播速度很快，但是持续时间较短。一般用户类似于社交圈子的朋友，用户之间的信息传播更多依靠亲戚朋友的口口相传的信任关系，这种传播方式速度较慢，但持续时间较长，在朋友圈子中更有影响力。



（a）超级用户



（b）次超级用户



（c）一般用户

图 5-27　不同类型用户在不同峰值时段转发路径长度比率分布

5.6 微博意见领袖识别

在新浪、腾讯、网易等微博平台上，一些经过实名认证的高级账户，即贵宾账户（VIP）拥有众多粉丝，粉丝数量通常达到几十万以上，这样的微博用户被网民称为“网络大 V”。网络大 V 相当于意见领袖，具有很大的影响力，他们的一次转发就会使得一条微博迅速火起来，成为网络热点话题，引导着网络舆论走向。要分析网络大 V 或意见领袖在网络舆情中所起的作用，首先需要解决微博意见领袖识别问题。

下面给出一种基于节点权重的微博意见领袖识别方法。

5.6.1 识别方法

基于节点权重的微博意见领袖识别方法的基本思路是根据网络拓扑特性，将网络抽象成一种有向图，通过分析节点之间结构关系，计算每个节点的权值，节点权值越大，成为意见领袖的可能性就越大。因此，可以将意见领袖识别问题归结为如何计算节点权重问题。

首先将微博网络抽象成一个有向网络图 $G = (V, E)$，其中 V 代表网络中的用户，称为节点；E 代表用户间的关系，称为连接节点之间的边。

由于每个用户拥有的朋友和粉丝数量不同，因此各个节点具有不同的权值，节点权值越大，说明该节点的影响力越大，成为意见领袖的可能性也就越大。在计算节点权重时，需要考虑到节点拥有的粉丝数量、节点连接关系以及交互关系等多种因素，以提高计算效率和精确度。

有效粉丝集合 Ef(u)定义如下：

$$\mathrm{Ef}(u) = \{v \mid v \in \mathrm{Follower}(u) \cap \mathrm{Response}(u) > \delta\} \tag{5-45}$$

式中，δ 是非负常数阈值，表示节点 u 的粉丝节点 v 对节点 u 反馈的门限，超过该阈值且属于节点 u 的粉丝的节点才能算作有效粉丝。

由连接关系所产生的节点权值 $IRL(u_i)$的计算方法如下：

$$\mathrm{IRL}(u_i) = \frac{\sigma}{N} + (1-\sigma) \sum_{u_j \in \mathrm{Follower}(u_i)} \frac{\mathrm{IRL}(u_i)}{L(u_i)} \tag{5-46}$$

式中，IRL(u_i)表示节点 u_i 连接关系产生的节点权值，Follower(u)为节点 u_i 所有粉丝集合，$L(u_j)$为节点 u_j 粉丝数目，σ 是介于 0 和 1 的阻尼系数，N 为网络图中的总节点数。

由节点交互关系所产生的节点权值 IRTR(u_i)的计算方法如下：

$$\mathrm{IRTR}(u_i)=\sum_{t_j\in\mathrm{Tweet}(u_i)}\frac{\sum_{u_j\in\mathrm{Response}(t_j)|}|Ns(u_j)-N\mu(u_j)}{|A|} \tag{5-47}$$

式中，IRTR(u_i)表示节点 u_i 的节点权值，Tweet(u_i)为用户 u_i 帖子集合，A 表示所有具有交互情况的帖子集|A|是 A 的集合，$Ns(u_j)$是节点 u_j 针对帖子 t_j 的响应次数，$N\mu(u_j)$为响应平均值，Response 包括用户转帖、回帖、评论和收藏。

节点综合权值 IR(u_i)的计算方法如下：

$$\mathrm{IR}(u_i)=(1-\beta)\times(\mathrm{IRL}(u_i)+\beta)\times\mathrm{IRTR}(u_i) \tag{5-48}$$

式中，参数β（$\beta\in[0,1]$）主要决定连接关系和节点交互关系在节点权值计算中所处的地位。当β较小时，节点权值主要由连接关系决定，特别当$\beta=0$ 时，则完全由连接关系计算权值。

综上所述，该方法的具体算法如下。

算法 5-5　基于多连接的节点权重算法（Multi-Link）

（1）利用网络爬虫工具，从互联网中采集实际的微博网络数据，提取其中的节点、连接等网络拓扑信息存入数据库待处理；

（2）构建有向网络图 $G=(V,E)$；

（3）利用公式（5-45）计算有效粉丝集合 Ef(u)；

（4）利用公式（5-46）计算由连接关系所产生的节点权值 IRL(u_i)；

（5）利用公式（5-47）计算由节点交互关系所产生的节点权值 IRTR(u_i)；

（6）利用公式（5-48）计算节点综合权值 IR(u_i)；

（7）计算网络图中所有节点的综合权值，并按综合权值由大到小排序，选取综合权值较大的 n 个节点，作为意见领袖的候选对象。

由于该方法在计算节点权重时考虑了节点粉丝数量、节点连接关系以及交互关系等多种因素，因此称为基于多连接（Multi-Link）的节点权重计算方法，它从计算效率和精确度两个方面改进了现有方法的不足，一方面，通过定义有效粉丝集合，将没有或拥有少量粉丝的节点排除掉，他们成为意见领袖的可能性极小，因为意见领袖或高权值节点必然拥有大量粉丝，这样就可大幅度减小网络图规模，有利于提高计算效率。另一方面，在计算节点权值时，不仅考虑了由粉丝产生的连接关系，还考虑了帖子的发布、转发、回复以及收藏等所产生的节点交互关系，因此提高了计算精确度。

5.6.2　算法验证

由于意见领袖的识别被量化成网络中节点权值序列，在这个序列中排名靠前的节点可认为是网络中的意见领袖。目前还没有用于衡量意见领袖识别效果的标准，主要采用算法对比方式来评价意见领袖识别效果。

下面通过实验数据对基于多连接（Multi-Link）算法和基于网络拓扑特性（Topological-based）算法的性能进行三种统计学方法的测试和对比，三种统计学方法包括 T-Test 检验、Kendall tau Rank 检验和 Spearman Rank 检验。

1．实验数据集

实验数据是从互联网中采集的真实社交网络数据，其数据集来源及规模如表 5-20 所示。

表 5-20　数据来源及规模

数据来源	帖子数目	用户节点
优酷	330232	5655
Youtube	4945382	1138499
其他论坛	53332256	1000000
新浪微博	2370238	350747

2．T-Test 检验

T-Test 检验也称 Student-t 检验，主要用于检验样本空间较小（例如 n<30）、总体标准差 σ 未知的正态分布数据。

首先使用 Multi-Link 算法和 Topological-Based 算法分别对 10 万个新浪微博用户节点进行意见领袖识别，得到前 100 位节点权值排名靠前的用户节点，然后对这 100 个用户节点使用 T-Test 检验，得到这些节点的 P-Value 分布。图 5-28 和图 5-29 分别给出了 Multi-Link 算法和 Topological-Based 算法的 T-Test 检验的 P-Value 分布。图中直线标识了 P-Value = 0.05 即 5%的分割线，可以看出，节点的 P-Value 值主要集中在该直线以下，即通过 T-Test 检验发现，两种算法计算的节点领袖权值具有较高可信度，能够代表网络中的意见领袖。

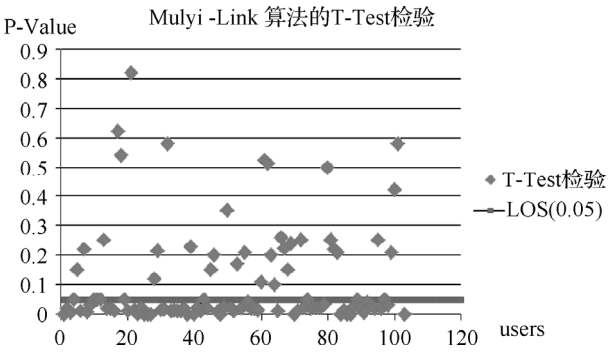


图 5-28　Multi-Link 算法的 T-Test 检验的 P-Value 分布

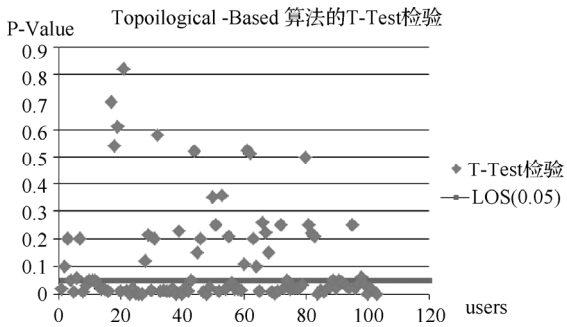


图 5-29　Topological 算法的 T-Test 检验

3．Kendall-tau 检验

在统计学中，肯德尔相关系数（Kendall-tau）是用来测量两个随机变量相关性的统计值，用 τ 表示其值。一个肯德尔检验是一个无参数假设检验，它使用计算得到的相关系数去检验两个随机变量的统计依赖性。τ 的取值范围在–1 到 1 之间，当 τ 为 1 时，表示两个随机变量拥有一致的等级相关性；当 τ 为–1 时，表示两个随机变量拥有完全相反的等级相关性；当 τ 为 0 时，表示两个随机变量是相互独立的。τ 的计算公式如下：

$$\tau = \frac{n_e - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{5-49}$$

（1）如果排列双方的排名是完美的（即两个排名是相同的），τ 值为 1；

（2）如果两个排列之间的分歧排名是完美的（即一个排名为扭转其他），τ 值为–1；

（3）对于所有其他 τ 值在–1 和 1 之间的排列，增加值意味着增加排列之间的排名。

根据计算结果，Multi-Link 算法和 Topological-Based 算法之间的 τ 值为 0.9107，说明这两种算法具有很高的一致性。

4．Spearman Rank 检验

在统计学中，斯皮尔曼等级相关系数（Spearman Rank）用来估计两个变量 X、Y 之间的相关性，其中变量间的相关性可以使用单调函数来描述，并用 ρ 表示其值。如果两个变量取值的两个集合中均不存在相同的两个元素，那么，当其中一个变量可以表示为另一个变量的很好的单调函数（即两个变量的变化趋势相同）时，两个变量之间的 ρ 值范围在–1 到 1 之间。

假设两个随机变量分别为 X、Y（也可以看作是两个集合），它们的元素个数均为 N，两个随机变量取的第 i（$1 \leqslant i \leqslant N$）个值分别用 X_i、Y_i 表示。对 X、Y 进行排序（同时为升序或降序），得到两个元素排行集合 x、y，其中元素 x_i、y_i 分别为 X_i 在 X 中的排行以及 Y_i 在 Y 中的排行。将集合 x、y 中的元素对应相减得到一个排行差分集合 d，其中

$d_i = x_i - y_i$，$1 \leqslant i \leqslant N$。随机变量 X、Y 之间的 ρ 值可以由 x、y 或者 d 计算得到，其计算方式如下：

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{5-50}$$

表 5-21 给出了 7 种算法之间的 Spareman Rank 值，从表 5-24 可以看出，Multi-Link 算法和 Topological-Based 算法具有较高的 Spareman Rank 值，序列一致性较高，说明 Multi-Link 算法和 Topological-Based 算法在意见领袖识别上表现出较好的能力。

表 5-21　各个算法的 Spareman Rank 值表

	A	B	C	D	E	F	G
A	—	0.9467	0.9528	0.8211	0.7731	0.8725	0.9214
B	0.9467	—	0.9235	0.7255	0.8421	0.6229	0.8756
C	0.9528	0.9235	—	0.8428	0.7815	0.7598	0.8437
D	0.8211	0.7255	0.8428	—	0.8542	0.8614	0.9102
E	0.7731	0.8421	0.7815	0.8542	—	0.9102	0.8415
F	0.8725	0.6229	0.7598	0.8614	0.9102	—	0.8910
G	0.9214	0.8756	0.8437	0.9102	0.8415	0.8910	—

注：各个字母所代表的算法，A:Topological；B:Topic；C:Multi-Link；D:PageRank；E:HITS；F:TwitterRank；G:InfluenceRank

5．算法性能对比

使用准确率和召回率评价意见领袖识别算法性能，准确率（P）和召回率（R）的计算公式如下：

$$P = \frac{A}{A+B},\ R = \frac{A}{A+C} \tag{5-51}$$

式中，A 为识别出的意见领袖数目，B 为识别出的非意见领袖数目，C 为未识别出的意见领袖数目。

由于在意见领袖识别中还没有标准来衡量是否发现全部的意见领袖，因此在计算准确率和召回率时通常以经验的意见领袖数目来近似真实的意见领袖数目。

表 5-22 为各种算法的召回率、准确度及平均节点处理时间。

表 5-22　各种算法的召回率、准确率及平均节点处理时间对照

算法	召回率（%）	准确率（%）	时间（min）/10w 节点
出度	57.31	62.24	0.1434
出度/入度结合	65.43	67.33	0.2355
ThreadRank	82.28	86.11	3.3765
InfluenceRank	81.71	84.70	2.8131
TwitterRank	88.53	90.48	2.7634
Topic-Based	90.73	92.54	2.8597
Topological-Based	91.22	92.36	2.2145
Multi-Link	89.38	91.75	1.3157

注：时间测试是在包含 10 万个用户节点的真实数据环境下得到的结果

由于 Multi-Link 算法采用微博网络拓扑结构中连接关系与节点交互相结合的计算方法，降低了网络节点规模，从而提高了计算速度，同时准确率和召回率也有显著的提高。图 5-30 和图 5-31 分别给出了不同算法的准确率和召回率以及计算时间比较。

从图 5-30 可以看出，在测试数据集上，Multi-Link、Topological-Based 及 Topic-Based 等算法的召回率和准确率比较高，与 TwitterRank 算法基本相当，比常见的出度和出度/入度结合算法更好，而出度和出度/入度结合算法的召回率和准确率都比较低。

从图 5-31 可以看出，出度和出度/入度结合两种算法的计算时间比较短，因为在计算过程中，这两种算法没有考虑其他的附加条件，算法比较简单，但召回率和准确率都比较低。而其他算法由于考虑了更多的修正因素，因此计算时间稍长。相比之下，Multi-Link 算法的计算时间处于中等水平。

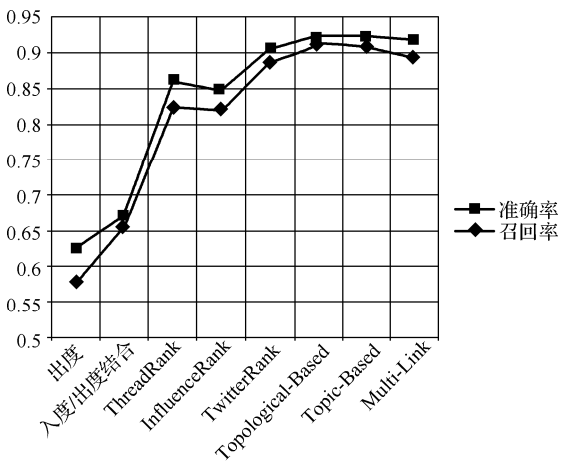


图 5-30　不同算法的准确率和召回率

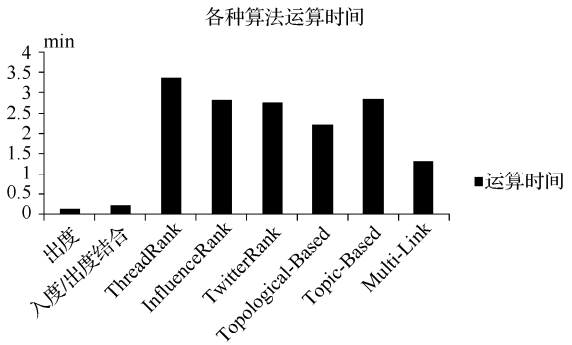


图 5-31　不同算法的计算时间比较

根据 T-Test、Kendall-tau 和 Spareman Rank 三种统计学检验方法的对比实验结果，表明 Multi-Link 算法具有较高的意见领袖识别能力，与 Topological-Based、Topic-Based 等算法具有一致性。

根据算法的准确率、召回率以及计算时间的实验结果，表明 Multi-Link 算法不仅在准确率和召回率上表现良好，并且比 Topological-Based、Topic-Based 等算法的计算时间要短，这对于处理海量网络数据来说是至关重要的。

因此，从意见领袖识别能力、准确率和召回率以及计算时间等综合指标来看，Multi-Link 算法更具优势。

参考文献

[1] S. Milgram. Behavioral study of obedience[J]. Journal of Abnormal and Social Psychology, vol. 67, no. 4, pp. 371-378, 1963.

[2] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities[C]. In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 160-168.

[3] A. Anagnostopoulos, R. Kumar, M. Mahdian. Influence and correlation in social networks[C]. In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, vol. 10, p. 7, 2008.

[4] S. Aral, L. Muchnik, A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks[J]. In Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 51, pp. 21 544-21 549, 2009.

[5] M. Gomez-Rodriguez, J. Leskovec, A. Krause, Inferring networks of diffusion and influence[C]. In

Proceeding of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 1019-1028.

[6] A. Agresti. Categorical Data Analysis[M]. 2nd, Wiley Series in Probability and Statistics, 2002.

[7] J. K. Benedetti, M. B. Brown. Strategies for the selection of log-linear models[J]. Biometrics, vol. 34, pp. 680–686, 1978.

[8] Zhilin Luo, Xintao Wu, Wandong Cai, Dong Peng. Examining Multi-factor Interactions in Microblogging based on Log-linear Modeling[C]. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, August 26-29, 2012, pp. 189-193.

[9] Ho T K. Random decision forests[C]. Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. IEEE, 1995, 1: 278-282.

[10] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1（1）: 81-106.

[11] Quinlan J R. C4. 5: programs for machine learning[M]. Morgan kaufmann, 1993.

[12] Breiman L, Friedman J H, Olshen R A, et al. Classification and regression trees[J]. 1998.

[13] Box G E P, Jenkins G M, Reinsel G C. Time series analysis: forecasting and control[M]. Wiley. com, 2013.

[14] Akaike, Hirotsugu. A new look at the statistical model identification[J]. IEEE Transactions on Automatic Control. 1974, 19 （6）: 716-723.

[15] Schwarz, Gideon E.. Estimating the dimension of a model[J]. Annals of Statistics, 1978, 6 （2）: 461-464.

[16] Unankard, Sayan, Chen, Ling, Li, Peng, Wang, Sen, Huang, Zi, Sharaf, Mohamed A., Li Xue. On the prediction of re-tweeting activities in social networks – a report on WISE 2012 Challenge[C]. In Proceedings of the 13th International Conference on Web Information Systems Engineering, Paphos, Cyprus, 28 - 30 November 2012.

[17] Palshikar G. Simple algorithms for peak detection in time-series[C]. In Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence. 2009.

第6章

网络论坛信息传播模型

6.1 引言

在 1.5 节中，对网络论坛及其信息传播模式进行了介绍。从中可以看出，网络论坛具有多元化、开放性、匿名性及互动性等特点，为网民提供了发表言论、获取信息的网络空间，网民通过发帖和回帖发表意见，参与观点传播和舆论形成，对于推进社会进步和政治民主起到了积极的作用，成为网络舆情的主要来源。同时，也容易被别有用心的组织和人员所利用，传播虚假消息和谣言，对人们的社会生活和意识形态造成负面影响。

由于网络论坛所具有的多元化、开放性、匿名性及互动性等特点，成为网络舆情的主要来源。因此，研究网络论坛中舆情形成规律，对于网络舆情识别与监测具有重要的意义。

本章主要对网络论坛观点传播与舆情形成、意见领袖识别、网络水军热帖检测、网络水军账号识别等问题进行分析和研究，有助于认识网络论坛观点传播与舆情形成的内在动力和规律。

6.2 网络论坛舆情形成模型

网络论坛具有多元化、开放性、匿名性及互动性等特点，成为网络舆论的主要阵地，极易演化成网络舆情。因此，研究网络论坛观点传播与舆情形成机制，对于网络舆论的引导和舆情的监测具有重要的现实意义。

国内外研究者对舆情形成模型进行了广泛研究，典型的模型是 Sznajd 模型[1]和 French-DeGroot 模型[2]，后来的很多模型都是这两种模型的扩展。Sznajd 模型是一个基于一维空间的舆情形成模型，认为节点是排列在一维空间上的点，并假定节点可以选择其中小数量的离散观点，由于该模型对选择过程给出了较好的解释而受到广泛关注，并被推广

到小世界网络和无标度网络。French-DeGroot 模型认为节点的观点可以在一个任意维度和结构的空间上延伸，观点在吸引力驱使下实时演化，即节点在观点空间中变换立场，趋向于其他节点观点比较集中的领域，该模型能够比较真实地反映现实网络中节点观点的倾向问题。

Sznajd 模型和 French-DeGroot 模型虽然在一定程度上表达了观点传播和舆情形成的主要特征，但也存在以下问题：一是 Sznajd 模型认为人们表达的观点只有两种：支持和反对，不能真实反映节点观点的模糊性和连续性；二是 French-DeGroot 模型假定网络中边的权值是不随时间变化的固定值，不能真实地反映节点间联系的亲密程度和时变特性。

下面给出一个基于节点影响力的网络舆情形成模型，在 French-DeGroot 模型的基础上，考虑到网络中节点观点变化的连续性以及节点间不同的连接强度及其时变特性，能够比较真实地反映网络论坛中的舆情形成过程。

6.2.1　网络论坛模型

在网络论坛中，用户之间通过发帖和回帖进行信息交互，其交互过程如下：首先原作者提出一个主题（thread），该主题有一个标题（title），并且在该主题中是唯一的。然后其他用户（也可以是原作者）围绕这个标题发一个或多个包含相应内容的帖子（post）展开讨论，图 6-1 为网络论坛主题组成图。

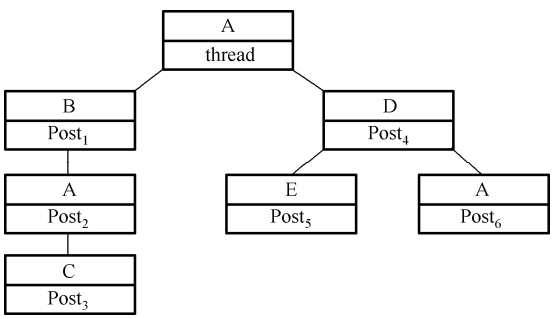


图 6-1　网络论坛主题组成图

为了描述网络论坛的观点传播和舆情形成过程，以网络论坛中的用户为节点、节点间相互连接为边来构建一个无向权值网络图，称为论坛网络，即：

$$G=(V, E, W) \tag{6-1}$$

式中，V 为论坛网络的节点数；E 为连接所有节点的边的集合，表示节点观点可能的传播路径；$W=[w_{ij}]$为边的权值矩阵，表示节点 i 对与之有连接的各个节点的影响力权值。

这里假设论坛网络是一个封闭的网络，舆情在论坛网络中产生，并仅在论坛网络中传

播。同时假设论坛网络中的节点数是不变的，每一个节点在 t 时刻都有一个用实数表示的观点值 $x(t)$。

6.2.2 舆情形成模型

在论坛网络中，节点间的连接因他们观点的不同具有不同的强度和内涵，并且是随时间动态变化的。为了描述这种特征，假设 t 时刻网络中节点 i 和 j 的观点分别为 $x_i(t)$，$x_j(t)$，则节点间的观点距离 d_{ij}^t 可表示为[3]：

$$d_{ij}^t = |x_i(t) - x_j(t)|, \forall i, j \in N, i \neq j \tag{6-2}$$

式中，d_{ij}^t 为节点 i 和节点 j 间的观点距离。

令

$$w_{ij}^t = \frac{1}{d_{ij}^t} = \frac{1}{\max(d, |x_i(t) - x_j(t)|)} \tag{6-3}$$

式中，d 是一个非常小的正数，以代替 $x_i(t)=x_j(t)$时分母为零的情况。公式（6-3）表明，两个节点间的观点距离越接近，彼此间的影响力就越大，这是符合实际情况的，例如社会中两个人的关系越亲密，当其中一个人遇到问题时愿意从亲密的朋友那里寻求建议和帮助，该朋友对他的影响力也就越大。

为了描述问题，定义影响力矩阵 $\boldsymbol{T}$ 为 $N \times N$ 的非负矩阵，则对所有的 $i, j \in N$，$T_{i,j}^t \in [0,1]$表示 t 时刻节点 i 对节点 j 的观点影响权重。同时 T 是一行随机矩阵，即 $\sum_{j=1}^{n} T_{ij}^t = 1$。对 $t \geqslant 0$，在 $t+1$时刻节点间的影响力可表示为：

$$T_{ij}^{t+1} = \frac{w_{ij}^t}{T_{ii}^t + \sum_{k \in N_{-i}} w_{ik}^t} \tag{6-4}$$

式中，T_{ii}^t 为 t 时刻节点 i 的自我影响力，N_{-i} 为除去节点 i 以外的节点集。

在 $t+1$时刻节点 i 的自我影响力可表示为：

$$T_{ii}^{t+1} = 1 - \sum_{k \in N_{-i}} T_{ij}^{t+1} \tag{6-5}$$

因此，一个舆情系统定义如下：

$$S = \{n, t, x(t), T, G_t(V, E, W)\} \tag{6-6}$$

式中，n 表示系统中的节点个数，$t=\{0, 1, 2, \cdots\}$为系统中离散的时间点，$x(t)=[x_1(t), x_2(t), \cdots, x_n(t)]$ 表示 t 时刻系统的观点剖面，$T=[T_{ij}]$ 是影响力矩阵，$G_t(V, E, W)$为 t 时刻系统的网络拓扑图。

舆情形成模型主要描述了论坛网络中用户观点随时间变化的趋势和网络舆情形成过程。设节点 $i \in V$ 在 $t \in T$ 时刻的观点为 $x_i(t) \in x(t)$，则节点 i 在 t+1 时刻的观点为 $x_i(t+1)$，可表示为[4]：

$$x_i(t+1) = \sum_{j=1}^{n} T_{ij}^{t+1} x_j(t) \tag{6-7}$$

式中，$x_i(t+1)$ 为节点 i 在 $t+1$ 时刻的观点值，T_{ij}^{t+1} 为节点 j 在 $t+1$ 时刻对节点 i 的影响力值。

舆情系统 S 在 $t+1$ 时刻的观点剖面可表示为：

$$x(t+1) = Tx(t) \tag{6-8}$$

式中，$x(t) = [x_1(t)x_2(t)...x_n(t)]^*$ 为 t 时刻的观点剖面，$T = [T_{ij}]$ 为 $N \times N$ 的影响力矩阵。

这样，在论坛网络中，每一节点随着时间的推移总是不断地改变着邻居节点的影响力并更新自身影响力。同时也会根据邻居节点当前的影响力和邻居节点上一时刻的观点来更新自身观点。节点观点演化的本质就是基于节点影响力的节点观点迭代过程。

为了进一步描述节点间的相互影响情况，引入 T 的 Laplacian 矩阵，T 的 Laplacian 矩阵定义为：

$$L = \mathrm{diag}(d) - A \tag{6-9}$$

式中，$\mathrm{diag}(d)$ 为 T 的对角矩阵，$d_i = \sum_{j \in V, j \neq i} T_{ij}$；$A$ 为 T 的邻接矩阵，$a_{ij} = \begin{cases} w_{ij}, & i \neq j \\ 0, & \text{其他} \end{cases}$。

这样，公式（6-8）可转化为：

$$x(t+1) = [I - L]x(t) \tag{6-10}$$

如果舆情系统 $S = [S[1], \cdots, S[n]]$ 中的任意两节点 i 和 j 均满足 $|x_i(t) - x_j(t)| < \varepsilon$（$\varepsilon$ 是一个很小的实数），则舆情系统 S 收敛。

可见，在论坛网络中，当 $t \to \infty$ 时，节点观点在外部环境的持续影响下不断发生改变。如果所有节点观点在某一时刻 t^* 均收敛于一定值 $S(t^*)$，则称该网络中的节点观点达到渐进一致或完全一致，形成网络舆情。

6.2.3　模型验证

下面通过仿真实验方法对论坛网络的舆情形成模型进行测试和验证。

1．仿真工具与步骤

在仿真实验方法中，采用 UCINET 6.0 软件作为论坛网络生成与分析工具，采用 Matlab 7.0 软件来仿真论坛网络的观点传播和舆情形成过程。

仿真步骤如下：

（1）初始化仿真起始时间 $t=0$，仿真时间 t_N。并通过 UCINET 6.0 随机生成包含 n 个节点的论坛网络，为各个节点随机分配一个初始观点值 $x_i(t)\in[1,5]$ 以及自我影响的初始影响力值 $T_{ii}^{t}\in[0,1]$。

（2）按照公式（6-4）和公式（6-5）分别计算每一节点在 $t=t+1$ 时刻对邻居节点及自身的影响力，构建影响力矩阵 T_{ij}^{t+1}（$d=10^{-2}$）。并按照公式（6-10）计算在 $t+1$ 时刻系统的观点剖面值。

（3）对任意节点 i 和 j，如果在 $t=t^*$ 时刻满足 $|x_i(t^*)-x_j(t^*)|<\varepsilon$（$\varepsilon=10^{-3}$），则结束仿真；否则，重复步骤（2）和步骤（3）。

2．模型有效性验证

为了验证模型的有效性，选取节点数 n 分别为 10、30、50 和 70 的 4 个不同规模的论坛网络进行仿真，得到如图 6-2 所示的舆情形成过程图，图中 Y 轴代表各节点的观点值，X 轴代表迭代次数。

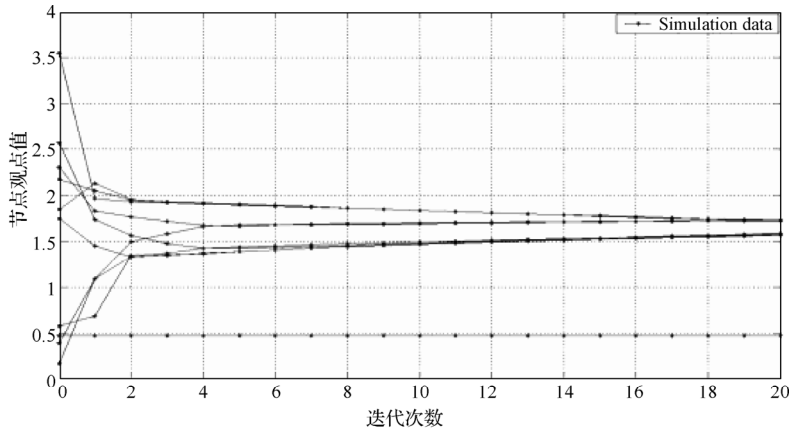
从图 6-2 显示的仿真结果可以看出：

（1）无论网络规模如何变化，模型均能收敛。即论坛网络中存在直接或间接连接的节点通过相互作用，各自持有的观点逐渐发生改变，最终收敛于某一定值。在这个过程中，观点传播在起始几个周期变化梯度较大，随着时间的推移变化趋缓，逐渐趋向一个稳定值，表明系统从非平衡态趋向平衡态。这与现实网络论坛中的观点传播趋势相一致，最初众人观点不一，随着彼此讨论、相互影响，最终观点趋于一致，形成舆情。

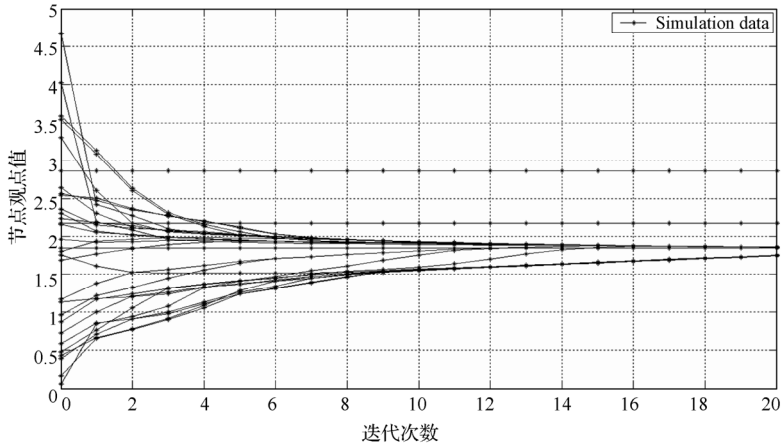
（2）观点最终形成两种不同的簇，即论坛网络中的节点最终形成两种持不同观点的群。这进一步验证了论坛网络具有的社区结构特性，即具有较高观点相似度值的节点可以划归于同一类型的社区，它们之间存在较多的连接；而社区与社区之间没有或有很少连接。

（3）对照网络图可以发现，通过较短的观点传播周期达到最终收敛值的节点存在更多的邻居节点。这与现实网络论坛中的情况相符，即邻居较多的用户因有更多的机会参与交互，其持有的观点更容易影响到其他用户，同时也更容易受到其他用户的影响，这些用户的观点状态趋于一致的速度更快。

（4）在节点数 $n=70$ 的论坛网络中，节点观点的收敛速度快于其他 3 个网络，说明网络规模并非是影响网络舆情形成的主要因素，网络规模大时，舆情形成速度可能更快。



（a）节点$n=10$下的舆情形成过程



（b）节点$n=30$下的舆情形成过程

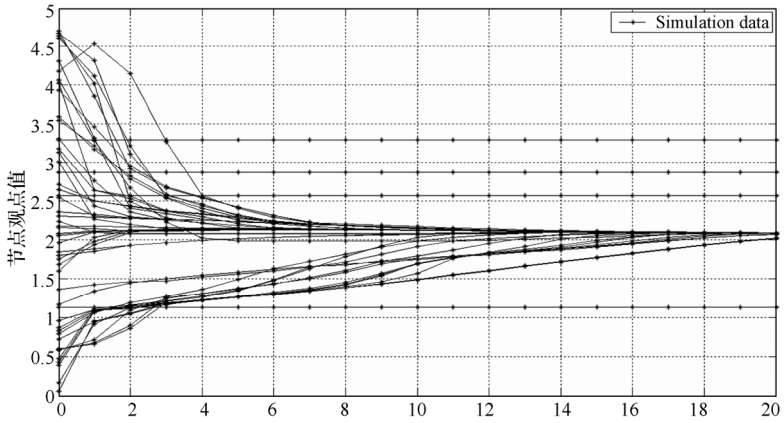
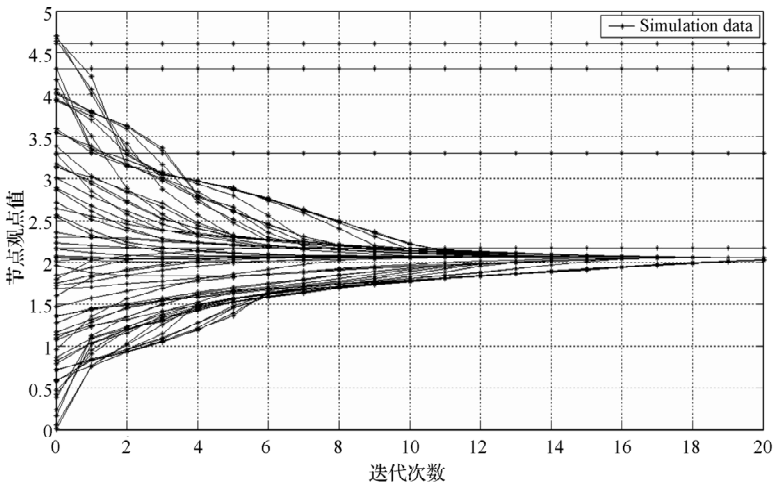


图 6-2　观点传播与舆情形成过程



（d）节点n = 70下的舆情形成过程

图 6-2（续）　观点传播与舆情形成过程

（5）孤立节点因不参与交互而不会影响到其他节点和受其他节点的影响，其观点状态不会发生改变（在图 6-2 中其变化趋势为直线）。在现实网络论坛中，这类节点对应于只浏览帖子而不发帖的用户。

6.3　网络论坛意见领袖识别

在网络舆论形成过程中，意见领袖发挥了积极的推动作用。在网络论坛中，大部分用户以浏览为主，对感兴趣的话题进行回复，他们的观点往往跟随意见领袖。在意见领袖的引导和影响下，局部观点或意见可能演化为网络舆情。因此，通过意见领袖来引导和控制网络舆情是十分重要的。要达到这一目的，首先需要解决网络论坛意见领袖识别问题。

在网络论坛意见领袖识别上，国内外研究者进行了广泛研究，提出了各种识别方法，包括简单统计测量法、影响力扩散模型、网页排名（PageRank）法、社交网络分析法等。其中，社交网络分析法采用社交网络分析中的相关量化指标来发现论坛中有影响力的用户。由于社交网络分析法主要关注静态网络的结构和统计学特性，很难反映出网络论坛中用户间交互关系的动态演变特性，对意见领袖识别的准确度产生一定的影响。

下面给出一种基于时间变化图的论坛意见领袖识别方法，将社交网络分析法和时间变化图相结合，提高了论坛意见领袖识别的准确度。

6.3.1　论坛有向网络图模型

将网络论坛的用户作为节点，一个用户对另一个用户的回帖看作是施加一个影响，如

B 对 A 进行了回复，则 B 对 A 施加了一个影响。这样，根据论坛用户间的交互关系，可以将网络论坛抽象成一个论坛有向网络图，如图 6-3 所示。在一个时间周期内，用户间就某一主题发出一定数量的帖子（包括发帖和回帖），则可将图 6-3 转换成以帖子数为边权值的论坛有向权值网络图，如图 6-4 所示。

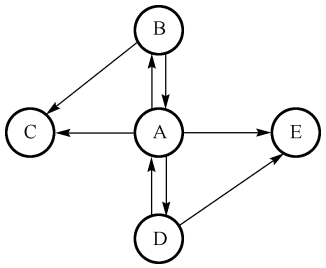


图 6-3　论坛有向网络图

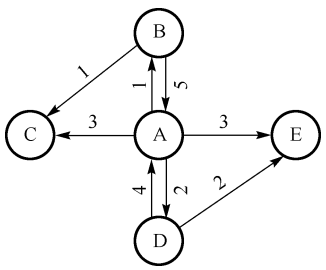


图 6-4　论坛有向权值网络图

事实上，基于论坛用户间交互关系的论坛网络是动态变化的，随着时间的推移，节点会不断地加入或离开网络，节点间的边会因此而发生变化，边的权值以及节点在网络中的角色和权限也会随之发生动态变化。图 6-5 为论坛网络动态演变过程，经过三个时间周期 T_1、T_2、T_3 的演化，形成一个有向权值网络图。

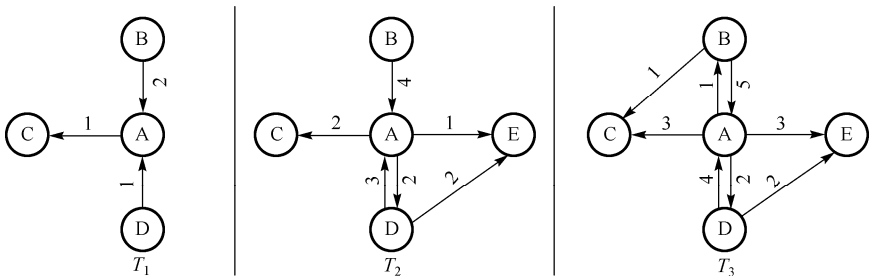


图 6-5　论坛网络动态演变过程

因此，基于论坛用户间交互关系的论坛网络图定义如下：

$$G = (V，E，W，T) \tag{6-11}$$

式中，V 为节点，表示网络论坛中的用户；E 为边，表示论坛用户间的交互关系；W 为边的权值，表示用户间交互的帖子数量；T 为网络图的生命周期，表示用户交互的持续时间。

生命周期 T可以分割成连续的子区间 $T = [t_0, t_1)，[t_1, t_2), \cdots, [t_k, t_{k+1})$，每一个区间$[t_k，t_{k+1})$用 T_k 表示，则在每一个 T_k 上的论坛网络图可以表示为 $G_k = (V[t_k, t_{k+1})，E[t_k, t_{k+1})，W[t_k, t_{k+1}), T_k)$，每一个时间窗口 T_k 对应的是论坛网络的一个即时快照。因此，一个论坛有向权值网络图可以表示为在各个时间窗口内的图序列[5]，即 $SF(T) = G_0, G_1, \cdots, G_k$。

6.3.2 论坛意见领袖识别算法

意见领袖是在观点形成与传播中扮演重要角色的人，他们在一个网络中的特殊位置以及交流习惯来影响其他用户的观点，给那些搜寻信息的用户提供一种导向，具有较大的影响力。因此，意见领袖识别问题可以转化为如何计算用户影响力，根据用户影响力值识别出意见领袖。

研究发现，网络论坛中的意见领袖通常具有以下两方面的行为特征：

（1）意见领袖总是与论坛中多个用户存在直接联系；

（2）在一定时间周期内，意见领袖往往非常频繁地直接与论坛中多个用户发生交互，并频繁向多个用户发布和回复帖子。

依据网络论坛中意见领袖的行为特征，可以采用节点度和聚类指标来量化论坛用户的影响力。

在论坛网络中，使用节点度来衡量一个用户直接与其他用户交流的频繁程度，节点度可以进一步分为出度（Out-degree）和入度（In-degree），分别表示一个用户在某一时间周期内发出和接收的帖子数量。如果一个节点具有一个高出度值，表示该用户在网络论坛中发出了比较多的帖子，因此有更多的机会去影响其他人；如果一个节点具有较高的入度值和非常低的出度值，表示该用户是一个很少参与交流的不活跃用户。出度和入度可以用如下公式表示：

$$\begin{aligned} D_o(i) &= \sum_{j \in N} \vec{\boldsymbol{e}}(i, j) w_{i,j} \\ D_I(i) &= \sum_{j \in N} \vec{\boldsymbol{e}}(j, i) w_{j,i} \end{aligned} \tag{6-12}$$

式中，i，j 分别代表图中的两个节点，$\vec{\boldsymbol{e}}(i, j) \in E$ 代表从节点 i 到节点 j 的一个有向边，$w_{i,j} \in W$ 代表有向边的权值，N 代表节点 i 的邻接节点集。

聚类是衡量一个用户与一个高度互联的用户群的亲密程度。在论坛网络中，节点的聚类又分为引入聚类（Incoming-clustering）和外出聚类（Outgoing-clustering），分别表示一个用户在某一时间周期内向用户群发出或接收的帖子数量。一个节点具有较高的外出聚类值，表示该用户发出的帖子可以快速地在用户群内传播，并可通过用户群传播到用户群以外的更大范围。因此，一个具有高外出聚类值的用户具有较大的机会成为意见领袖。同理，如果一个节点具有较高的引入聚类值，表示该节点与较多的用户群存在连接，信息来源较多，有更多的机会来接受他人的意见。

外出聚类和引入聚类值可以用如下公式表示：

$$C_o(i)=\frac{\sum_{j\in N}D_o(j)}{D_o(i)\times(D_o(i)-1)}$$
$$C_I(i)=\frac{\sum_{j\in N}D_I(j)}{D_I(i)\times(D_I(i)-1)} \tag{6-13}$$

式中，j 表示用户群中的节点，$\sum_{j\in N}D_o(j)$ 和 $\sum_{j\in N}D_I(j)$ 表示用户群中节点间实际存在的边。

通过以上分析可以得出：具有高出度值和高外出聚类值的用户具有很大的影响力，在某一时间窗口内成为意见领袖的可能性很大。另外，一个具有高出度且入度为 0 的用户很可能是一个恶意的信息发布者。因此，可以采用如下公式来量化用户的影响力：

$$\text{Influence}(i)\ =\ \tanh(D_o(i))\cdot(\alpha D_o(i)+\beta C_o(i)) \tag{6-14}$$

式中，α、β 为加权值；$\tanh(D_o(i))$表示出度的双曲正切值，表示当出度等于或接近 0 时，用户最终的影响力值等于或接近于 0。

在确定 $D_o(i)$和 $C_o(i)$之前，加权值α、β是未知的。可以采用 AHP 方法来求解，根据 Saaty“1-9”标度法构造判断矩阵：

$$R=\begin{bmatrix}A_1/A_1 & A_1/A_2\\ A_2/A_1 & A_2/A_2\end{bmatrix}=\begin{bmatrix}1 & 3\\ 1/3 & 1\end{bmatrix}$$

分别计算两行元素的几何平均值，并作归一化处理，计算最后的权重值，计算公式为：

$$W_i=\frac{W_i'}{\sum_{i=1,\cdots,n}W_i'} \tag{6-15}$$

式中，W_i为权重值，W_i'为判断矩阵中每行元素的几何平均值。

α、β的计算结果为：

$$\alpha=\frac{\sqrt{1\times 3}}{\sqrt{1\times 3}+\sqrt{1\times\frac{1}{3}}}=\frac{1.7321}{1.7321+0.5774}\approx 0.75$$

$$\beta=\frac{\sqrt{1\times\frac{1}{3}}}{\sqrt{1\times 3}+\sqrt{1\times\frac{1}{3}}}=\frac{0.5774}{1.7321+0.5774}\approx 0.25$$

使用公式（6-14）分别计算每个时间窗口 T_k 内的节点影响力，选取前 n 个影响力值最大的用户组成集合，并对每个时间窗口内的结果进行匹配，即可跟踪和识别出随时间演变的网络论坛意见领袖。

6.3.3 算法验证

下面通过实验数据对网络论坛意见领袖识别算法性能进行测试和验证。

1．实验数据集

实验数据来源于新浪网财经论坛（http://club.finance.sina.com.cn）的技术交流版块，通过网络爬虫工具获取了 2011 年 4 月～10 月间的发帖数据。该时间区间共有 5128 个帖子和 421 个参与发帖的用户。

按照以下规则构建论坛网络图：

（1）如果发帖人对自己所发的帖子进行回复，则不建立节点的自我指向边；

（2）如果发帖人的帖子无回帖或只有自己回复，则删除该节点；

（3）如果回帖人 B 对发帖人 A 的帖子进行了回复，则认为 B 对 A 施加了一个影响。即在两个节点间建立由 B 指向 A 的边，边的权值根据回复的次数而定。

2．算法有效性验证

为了验证算法的有效性，将算法所获取到的结果与静态网络图（以连续 5 个月为时间窗口）中所获取到的结果进行相似度值计算和对比。

首先以连续 5 个月为时间窗口构建论坛静态网络图，如图 6-6 所示，依据公式（6-14）计算出静态网络中的前 20 个意见领袖。然后使用本算法，以 4 月 1 日为起始点，以 15 天的时间周期为时间步长，构建出具有 12 个不同时间窗口的网络图。

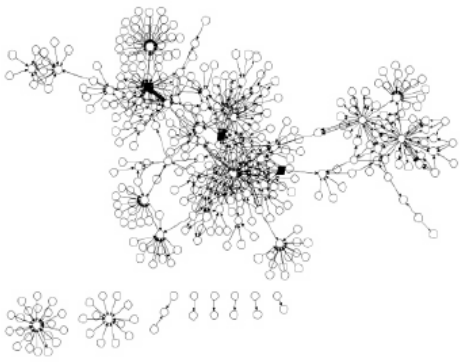


图 6-6　论坛静态网络图

为了选取算法中一个最优的 TOP n，分别获取每个图中前 10 个和前 20 个潜在的意见领袖。将获取的潜在意见领袖集合分别与图 6-6 中识别出的意见领袖集合进行相似度计算，具体采用 Jaccard 系数来度量相似度，因此需要计算它们之间的 Jaccard 系数，建立对应于 TOP 10 和 TOP 20 的相似度值分布图，如图 6-7 和图 6-8 所示。

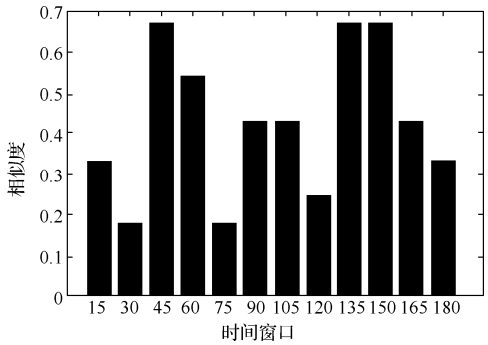


图 6-7　TOP 10 潜在意见领袖相似度值

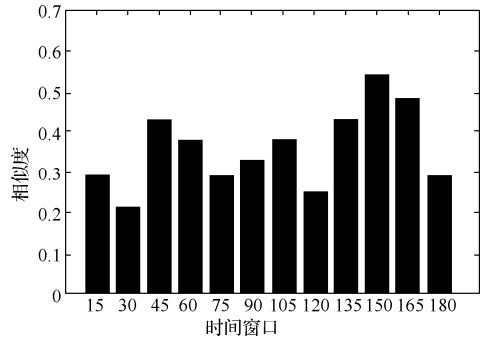


图 6-8　TOP 20 潜在意见领袖相似度值

从图 6-7 和图 6-8 可以看出，两个图的相似度值分布总体上比较相似，但图 6-7 相似度值平均要比图 6-8 高出 16%，说明采用前者的识别准确率更高。表 6-1 是结合公式（6-14）和图 6-6 计算出的影响力值排名前 10 的用户列表。

依据 TOP n =10 选取算法对 12 个潜在意见领袖组成的集合相互进行匹配，所获取的用户分别为：历尽风雨见彩虹、云天梦、俺 Q1395278391，并将以上 12 个集合与表 6-1 中的结果进行 Jaccard 系数计算，得到如图 6-9 所示的相似度变化图。

从图 6-9 可以看出，不仅各个时间步长上的相似度值低于 1，相邻步长的相似度值垂直变化也比较快。

表 6-1　影响力值排名前 10 的用户列表

UserId	Outdegree（出度）	Outgoing-clustering（外出聚类值）	Influence-value（影响力值）
历尽风雨见彩虹	59	0.98	44.50
财经小数	52	0.85	39.21
fedsrgggggggg	50	0.78	37.70
frgtgt	48	0.74	36.19
云天梦	45	0.89	33.97
渐行渐远渐无言	41	0.75	30.94
俺 Q1395278391	40	0.87	30.22
飘飞的雪泥	38	0.72	28.68
2410897114hdd	37	0.72	27.93
伊凡童心	22	0.43	16.61

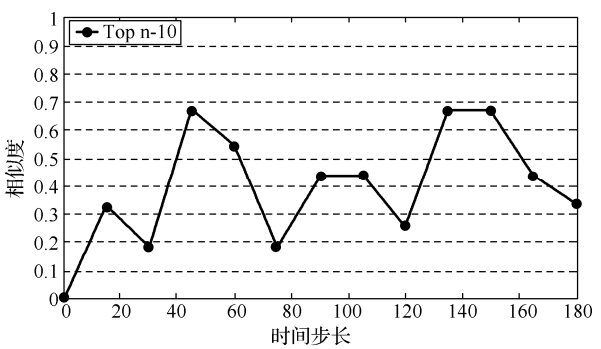


图 6-9　相似度变化图

实验结果表明，根据网络论坛中用户影响力随着时间动态变化这一特性，将时间周期分割成不同的时间窗口，分别计算每个时间窗口的节点影响力，能够有效地提高意见领袖识别的准确率。

6.4　网络论坛水军热帖检测

由于网络论坛具有多元化、开放性、匿名性及互动性，成为广大网民发表言论、获取信息的重要网络平台，也是网络舆情形成的主要网络空间。网络舆情包括正负两个方面，正面网络舆情是由网民发帖、点击和回帖形成的网络舆情，反映了公众对现实生活中的某些热点、焦点问题所持的具有较大影响力和倾向性的言论和观点。负面网络舆情主要是由造谣者撒布的网络谣言或者由网络水军炒作而引发的虚假网络舆情，对人们的社会生活和意识形态造成负面的影响。因此，网络水军炒作行为是引发虚假网络舆情的主要来源和推动力。

网络论坛水军（简称网络水军）炒作行为是指网络水军就某个话题发帖和回帖，推动该话题迅速形成网络论坛热点话题，引起广大网民的关注，进而引发虚假网络舆情。可见，这种热点话题是由网络水军通过发帖和回帖推动的，因此称为网络水军热点话题，其帖子称为网络水军热帖。

通常，网络论坛热点话题从产生到消失需要经历一个包括潜伏期、显现期、演进期、衰退期、消解期等五个阶段[6]的生命周期，如图 6-10 所示。在这些阶段，热点话题和一般话题一样，也有一个发生、发展和消失的过程，也就是从量变到质变的过程。在热点话题发生前，总会有一些征兆出现。只要及时捕捉到这些信息，加以分析处理，就能及时检测到话题幕后的推动力量，并对话题的演化过程有一个基本的认识，从而采取必要的应对措施。

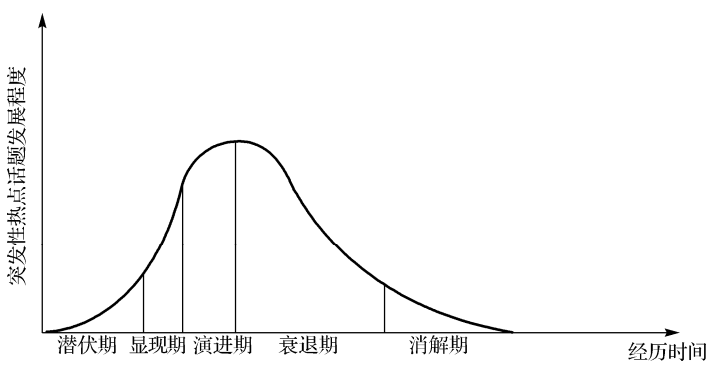


图 6-10　热点话题的生命周期图

网络水军对舆论的引导过程分为主题吸引和观点引导两个阶段。首先，他们将炒作的话题信息密集发布于各个网络论坛上，并通过在短时间内大量的回帖来吸引网民的眼球，引发公众围观效应。然后通过角色扮演与情感“认同”，有理有据的逻辑表达和团体协同

生产等手段进一步影响网民的思想，达到左右舆论的目的。主题吸引阶段对应于热点话题生命周期中的潜伏期和演进期前期，在这一阶段，主要由网络水军不断地发帖、回帖，网民还很少参与其中。一旦进入到观点引导阶段，大量的网民参与其中，网络水军的作用就不明显了。因此，网络水军热帖检测应侧重于潜伏期。

6.4.1 热点话题特征提取

对于网络水军热帖检测，首先需要分析网络水军推动的热点话题或热帖在潜伏期内的基本特征，定义并提取热点话题特征参数。然后采用机器学习算法对网络水军热帖进行分类检测，从网络论坛热点话题中识别出网络水军热帖。

一个网络论坛热点话题可以用如下四元组来定义：

$$X=\{H_l,R_l,S_l,D_l\} \tag{6-16}$$

式中，H_l 为热点话题潜伏期回帖指数，R_l 为热点话题潜伏期新注册 ID 指数，S_l 为热点话题潜伏期简单回帖指数，D_l 为热点话题潜伏期用户 ID 离散指数。

回帖指数反映了用户针对该话题的回帖数增幅情况，定义如下：

$$H_l=\lg\frac{\sum_{i=0}^{t}p_i}{t\lambda} \tag{6-17}$$

式中，p_i 为在话题潜伏期内某一时刻 i 的回帖数，t 为话题潜伏期时间，λ 为在话题潜伏期内正常回帖率阈值。该指数考虑了由网络水军推动话题时，回帖数会呈现“指数”级的增长而非线性增长[5]。

新注册 ID 指数反映了新注册的用户 ID 占该时期所有用户 ID 的比例，定义如下：

$$R_l=\frac{r}{R} \tag{6-18}$$

式中，r 为话题潜伏期内新注册的用户 ID 数，R 为话题潜伏期内所有用户 ID 数。

简单回帖指数反映了在话题潜伏期内内容简单的回帖所占的比例，定义如下：

$$S_l=\frac{s}{\sum_{i=0}^{t}p_i} \tag{6-19}$$

式中，s 为在话题潜伏期内简单回帖数。该指数考虑到在潜伏期有限的时间内，网络水军为使话题尽快演变成热点话题，往往注重回帖的数量而非回帖本身的质量，所以

其回帖内容经常比较简单，包含的字符数也比较少。p_i 为在话题潜伏期内某一时刻 i 的回帖数。

用户 ID 离散指数反映了在话题潜伏期内同一用户 ID 发多条回帖的数量占该时期所有用户 ID 的比例，定义如下：

$$D_l = \frac{d}{R} \tag{6-20}$$

式中，d 为话题潜伏期内同一用户 ID 回帖数大于阈值 η 的个数，R 为话题潜伏期内所有用户的 ID 数。该指数考虑到网络水军为使话题尽快演变成热点话题，往往在话题潜伏期内利用同一用户 ID 发表多条回帖。如果该时期同一用户 ID 的回帖数大于 η（在话题潜伏期内正常状况下同一用户 ID 回帖数阈值），则用户 ID 有网络水军注册的嫌疑。该指数还考虑到网络水军为了隐藏自己的身份，突破网络论坛对一个用户 ID“单日发帖数量限制”的约束，将会注册多个用户 ID 进行发帖、回帖。

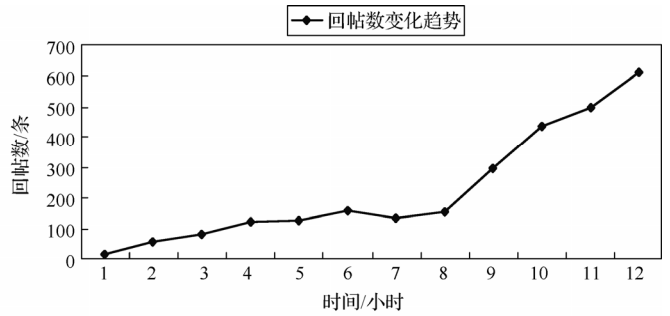
通过对事后确认由网络推手组织、网络水军推动的热点话题的研究，发现在此类话题的演变过程中，话题热度虽然变化较大，但在回帖指数、简单回帖指数、新注册 ID 指数、用户 ID 离散指数这 4 个特征参数上均有明显的规律可循。例如，2010 年 10 月 5 日天涯社区娱乐八卦版一篇名为《感谢这样一个极品的朋友给我带来这样一个悲情的国庆》（后简称“小月月”事件）的水军帖，该帖产生后的前 12 个小时内，共有 1505 个用户 ID 参与其中，回复帖 2707 条。图 6-11 统计了该话题的 4 个特征参数在此时间区间上的变化情况。图中横坐标表示时间，纵坐标表示某一特征参数在对应时间区间上的大小。

从图 6-11 可以看出：

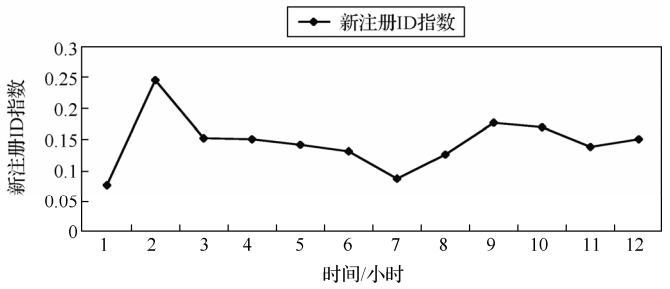
（1）图 6-11（a）显示，在开始的前 3～4 个小时内，在网络水军的推动下，该话题的回帖数量整体处于上升趋势但变化幅度不大。随着不断演化，话题逐渐显现，从第 8 个小时开始，话题从显示期完全过渡到演进期，大量的网民参与其中，回帖数量急剧上升。

（2）图 6-11（b）和图 6-11（c）显示，话题的新注册 ID 指数和简单回帖指数在前 2 个小时都比较高，随着时间的推移不断下降，从第 3 个小时开始，数值分别保持在 0.15 和 0.25 左右，这与由网络水军推动的热点话题所呈现的特征相吻合。处于潜伏期和显示期阶段的话题，主要是由于网络水军参与其中，网络水军通过注册新的 ID 并发送大量的简短回帖来推动话题，后期随着参与回帖的普通网民人数的不断增加，新注册用户 ID 和简单回帖数量所占的比例较潜伏期和显示期有所下降，网络水军的作用变弱。

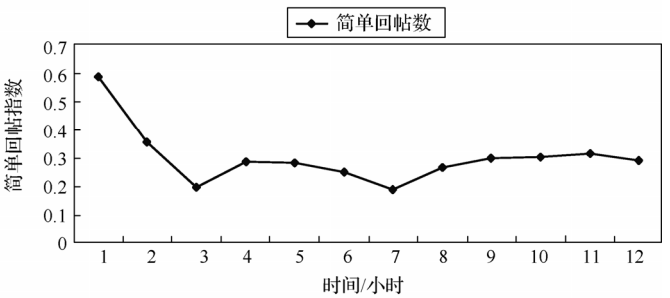
（3）图 6-11（d）显示，随着话题过渡到演进期，大量的网民参与其中，同一用户 ID 发多条回帖的数量（这里对一个小时内发三条以上的帖子的用户 ID 进行标示）占所有用户 ID 数量的比例不仅较话题潜伏期和演进期有所降低，而且变化趋缓。



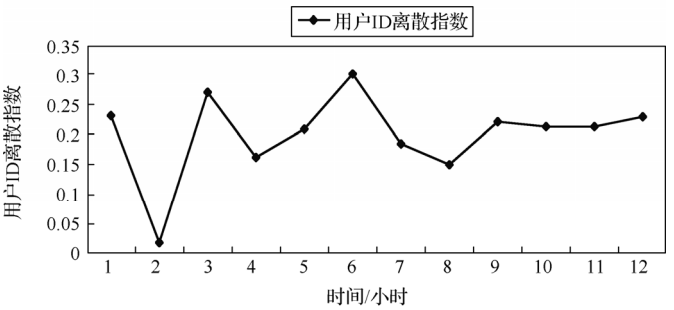
（a）回帖变化趋势图



（b）新注册ID指数变化趋势图



（c）简单回帖指数变化趋势图



（d）用户ID离散指数变化趋势图

图 6-11　网络水军热帖统计特征

6.4.2　水军热帖检测算法

在网络水军热帖检测中，采用支持向量机 SVM（Support Vector Machine）学习算法。SVM 是一种经典的机器学习方法，以统计学习理论为基础，对于小样本学习问题，表现出很强的认知能力。SVM 的基本思想是在二维两类线性可分情况下，有很多可能的线性分类器将一组数据分割开，但是只有一个使两类的分类间隔最大。SVM 分类就是寻找一个最优分类面，尽量使该平面能够满足分类的限制条件，可以把需要分类数据集合中的所有点分开，并且尽可能地使点与该分类面距离最远。

网络水军热帖检测方法的基本思想是通过定义和提取热点话题的特征参数，利用 SVM 分类器对热点话题幕后推动力进行分类，识别出热点话题是人为推动的还是自然形成的，如果是人为推动的，则是网络水军热帖。

网络水军热帖检测方法形式化描述如下：

给定训练集$\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$，其中 $x_i \in X$，X 表示输入空间，是所有热点话题的集合；$y_i \in Y = \{-1, +1\}$，Y 表示输出域。当热点话题由人为推动时，$y = 1$；自然形成时，$y = -1$，m 为样本数目，$1 \leqslant i \leqslant m$。

显然，网络水军热帖检测是一个二分类问题。根据上述的定义，基于 SVM 的网络水军热帖检测方法就是设计一个最优分类函数 $f(x): X \to Y$，使它能够找到一个最优的线性分类面，不仅能把正常热点话题和异常热点话题分离开，还要使分类间隔最大。由于分类面函数可以表示为：

$$g(x) = w \cdot x + b \tag{6-21}$$

式中，w 为超平面的法向量，b 为超平面的偏移向量。

求解最优分类面的过程可以用如下最优化问题描述：

$$\min \frac{1}{2} = \|w\|^2 \tag{6-22}$$

这是一个二次规划问题，其最优解可由如下 Lagrange 函数的鞍点给出：

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \{y_i (w \cdot x_i + b) - 1\} \tag{6-23}$$

式中，α_i 为 Lagrange 乘子，$\alpha_i \geqslant 0$。

问题变为对 w 和 b 求 $\min L(w, b, \alpha)$，可以把原问题转化为如下较简单的对偶问题：

$$\max Q(a)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,l}^{m}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{6-24}$$

求解上述问题，得到如下最优分类函数：

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{m}\alpha_i y_i(x\cdot x_i)+b\right) \tag{6-25}$$

整个算法实现分为两个阶段：第 1 阶段是学习阶段，采用 SVM 学习算法求解满足检测精度的 SVM 最优分类函数 f 和最优训练集 T；第 2 阶段是检测阶段，针对待检测热点话题样本集合 X，使用最优分类函数 f 按照顺序对集合中样本进行检测。

6.4.3 算法验证

下面通过实验数据对网络水军热帖检测算法性能进行测试和验证。

1．实验数据集

网络论坛中的热点话题一般由众多主题相似的帖子组成，并且这些帖子分布在多个论坛中，获取不同热点话题在多个论坛中的所有帖子是一件非常困难的事情。因此，实验以单一的网络论坛作为数据源，通过网络爬虫工具获取了网易新闻论坛（http://bbs.news.163.com）2011 年 3 月 1 日到 2011 年 5 月 1 日间的数据。这一时间区间共有 4248 条帖子，包含 2716 个话题，并且有超过 2000 条帖子是孤立的。通过热点话题发现算法提取出 5 个热点话题，包含 9 条具有时间突发特性的热帖，并按以下方法对获取的数据进行处理：

（1）选取潜伏期 $t=4$ 小时，即只保留主帖自创建开始 4 小时内的回帖数据，包含回帖用户 ID、回帖时间、回帖内容、回帖用户 ID 注册时间；

（2）选取 $\lambda=50$ 条/小时，即主帖在潜伏期内的回帖率低于 50 条/小时是正常的；

（3）由于人为推动的热点话题并不需要太长的时间，所以选取新注册用户 ID 的时间阈值为 3 个月，基本包括了网络水军注册的绝大部分用户 ID，即用户 ID 注册时间少于 3 个月的视为新注册 ID；

（4）选取简单回帖内容的阈值为 10 个字符，即除去回帖内容中包含的图片、表情和标点符号后，字符个数少于 10 个的回帖视为简单回复。如果回帖内容中包含很多重复部分，重复部分少于 10 个字符，也视为简单回复；

（5）选取 $\eta=3$，即同一用户 ID 在潜伏期内对同一主帖的回帖数超过 3 条，视为该用户 ID 有网络水军注册的嫌疑。

依据以上数据处理方案，并通过公式（6-17）到公式（6-20）计算，可以得到如表 6-2 所示的 9 条热帖对应的量化参数。

由于实际网络论坛中热点话题幕后推动力存在异常（网络水军推动）的概率低，数量少，异常的规模难以刻画，例如在表 6-2 的 9 条热帖中，只有 5 号主帖存在明显的人为推动迹象。因此实验数据是由自然形成的热点话题与网络水军推动的热点话题构成的合成话题。同时，为了更好地考察 SVM 主动学习算法的泛化能力，使实验数据保持一定的规模，在上述 9 条热帖的基础上人为加入“小月月”事件、“凤姐”事件、“封杀王老吉”事件等 7 条已被证明是由网络水军推动的热帖，使样本集 X 中的正负样本比例为 1:1。网络水军热帖量化参数如表 6-3 所示。

表 6-2　热帖对应的量化参数

序号	主帖标题	回帖指数	新注册 ID 指数	简单回帖指数	ID 离散指数
1	中石油给日本捐款 3000 万	–1.30	0.56	0.10	0.11
2	为药家鑫的判决赌一把	–0.28	0.27	0.03	0.17
3	你做法官，如何判药家鑫？	–0.75	0.17	0.14	0.13
4	拒绝死刑，挽救药家鑫，现征集万民签名	–1.60	0.67	0.10	0.10
5	新浪微博无辜封杀一剑传媒草根团队微博	0.41	0.69	0.38	0.39
6	蹲监“被死亡”，千万资产乡领导贱卖据已有	–1.35	0	0.11	0.33
7	欢呼吧！药家鑫被判死刑！	–0.40	0.14	0.03	0.57
8	震惊！副乡长抢夺民企谁汗颜？	–1.07	1	0.06	0
9	我理解地平线网友对药家鑫案的观点	–0.61	0.2	0.04	0.50

表 6-3　网络水军热帖量化参数

序号	事件标题	回帖指数	新注册 ID 指数	简单回帖指数	ID 离散指数	出处
1	“小月月”事件	0.15	0.15	0.49	0.15	天涯社区
2	“凤姐”事件	–0.11	0.24	0.46	0.05	天涯社区
3	封杀“王老吉”事件	0.81	0.30	0.51	0.04	天涯社区
4	“康师傅”水源门事件	0.21	0.27	0.38	0.23	天涯社区
5	“犀利哥”事件	0.49	0.33	0.22	0.32	天涯社区
6	“奥巴马女郎”事件	0.64	0.36	0.35	0.16	猫扑社区
7	“贾君鹏”事件	1.16	0.60	0.79	0.42	魔兽世界吧

2．算法性能对比

在保证表 6-2 中实验数据不变的情况下，分别利用 SVM 主动学习算法与文献[7]所采用的综合指标法进行热点话题分类，评价指标是准确率、召回率和 F1 值，两种方法的检测效果对比如图 6-12 所示。从图 6-12 可以看出，相比于综合指标法，SVM 主动学习算法的检测效果更好，平均准确率达到 80%以上。

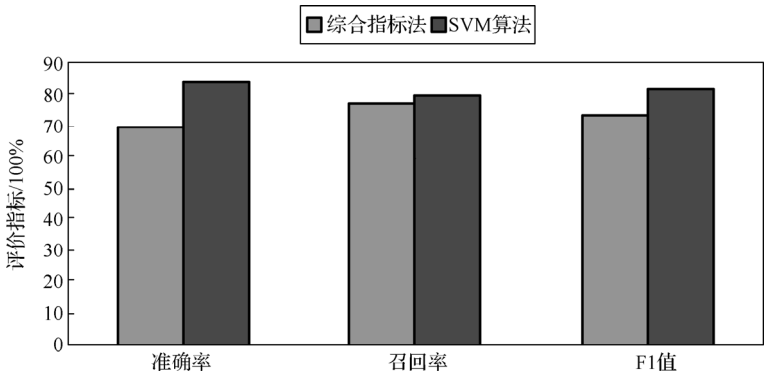


图 6-12　SVM 主动学习算法和综合指标法检测效果对比

6.5　网络水军账号检测

网络水军受雇于网络公关公司，通过为他人发帖、回帖、造势来获得报酬，他们利用大众惯用的沟通方法在论坛、社交网站等平台以聊天方式为个人或公司做宣传或攻击，通过文章和评论来试图达到影响、引导和制造网络舆论的目的。

在网络水军活动中，通常包括三类主体：客户、网络公关公司和网络水军，网络公关公司收到客户委托后，作为任务下发给网络推手，网络推手组织网络水军完成其任务。这样，网络公关公司、网络推手、网络水军就形成了灰色利益链，他们在实现客户目标的同时也获得了自身利益。

根据客户目标的不同，网络水军的任务一般分为两类：广告宣传和网络炒作。第一类任务是通过增加指定内容的可见率达到广告宣传的目的；第二类任务则是通过炮制网络热点，吸引广大网民围观和讨论，达到网络炒作的目的。为了完成第一类任务，网络水军需要以最快速度在各种尚没有出现该信息的网络论坛以主帖的形式发表指定内容，使其在最短时间内扩散。为了完成第二类任务，网络水军则需要短时间内在各大网络论坛大量发帖、回帖，使炒作对象在网络论坛长时间处于显眼位置，吸引网民关注，引发讨论，形成网络热点。为了完成网络炒作任务，网络水军会在全国各大论坛注册多个账号（也称为网络马甲），以不同身份登录论坛，针对论坛上若干主帖在短时间内大量回帖，达到网络炒作的目的。

下面介绍一种网络炒作的网络水军账号检测算法[8]。

6.5.1　检测算法

1．算法基本思想

算法采用“层层逼近，逐步求精”的策略，采用人类行为统计分析、社会网络结构分

析、时间特征分析等方法逐步排除正常用户和数据，不断缩小计算范围，最终确定网络水军账号。

算法流程如下：

（1）首先采用人类行为统计分析方法，统计论坛单日回帖数、日人均回帖数和日帖均回复数，将不可能发生网络炒作的时段排除，提炼出可疑区间，只对可疑区间做下一步分析，缩小分析范围；

（2）然后采用社会网络结构分析方法，对可疑区间构建单日用户协作网络，排除没有发生大规模用户协作现象的时段，提炼出高可疑数据，只对高可疑数据做下一步分析，进一步缩小分析范围；

（3）最后采用时间特征分析方法，对高可疑数据的用户回复行为时间特性进行分析，最终判定是否为网络水军。

2．论坛可疑时段识别

网络论坛的单日回帖数服从幂律分布，即大部分时间的论坛单日回帖数很小，而少数日子论坛单日回帖数很大。为了制造轰动效应，达到网络炒作的目的，网络水军通常使用多个账号在短时间内针对网络论坛上若干主帖大量地回帖，导致网络论坛的单日回帖数、日人均回帖数和日帖均回复数明显增大。因此，可以将网络论坛的单日回帖数、日人均回帖数和日帖均回复数作为识别可疑时段的指标，如果某个时段的这三项指标都大于均值，则确定该时段为可疑时段。

（1）论坛单日回帖数。该指标定义为论坛 t 日提交的回帖数之和，记作 RN_t，则有：

$$\mathrm{RN}_t = \sum_{u \in U_t} \mathrm{RN}_u^t \tag{6-26}$$

式中，N_t 为 t 日提交过回帖的用户集合，RN_u^t 为用户 u 在 t 日的回帖数。将单日回帖数大于等于均值的时段记作 S_1，则有：

$$S_1 = \left\{ t, \mathrm{RN}_t \geqslant \frac{\sum_{t \in T} \mathrm{RN}_t}{|T|} \right\} \tag{6-27}$$

式中，T 为数据集涵盖的时段，$|T|$为数据集包含的天数，下同。

（2）论坛日人均回帖数。该指标定义为论坛 t 日回帖数与当天提交过回帖的用户数之比，记作 ARNU_t，则有：

$$\mathrm{ARNU}_t = \frac{\mathrm{RN}_t}{|U_t|} \tag{6-28}$$

将日人均回帖数大于等于均值的时段记作 S_2，则有：

$$S_2=\left\{t,\mathrm{ARNU}_t\geqslant\frac{\sum_{t\in T}\mathrm{ARNU}_t}{|T|}\right\} \tag{6-29}$$

（3）论坛日帖均回复数。该指标定义为论坛 t 日回复数与当天被回复过的主帖数之比，记作 ARNP_t，则有：

$$\mathrm{ARNP}_t=\frac{\mathrm{RN}_t}{|P_t|} \tag{6-30}$$

式中，P_t 为当天被回复过的不同主帖的集合，将日帖均回复数大于等于均值的时段记作 S_3，则有：

$$S_3=\left\{t,\mathrm{ARNP}_t\geqslant\frac{\sum_{t\in T}\mathrm{ARNP}_t}{|T|}\right\} \tag{6-31}$$

将单日回帖数、日人均回帖数、日帖均回复数均大于均值的时段定义为论坛可疑时段，记作 S，则有：

$$S=S_1\cap S_2\cap S_3 \tag{6-32}$$

3．用户单日回复模式分析

排除不可能发生网络炒作的时段后，采用构建用户协作网络的方法对可疑时段的用户单日回复模式进行分析。

（1）用户协作性描述

为达到网络炒作的目的，网络水军必定会使用多个账号短时间内针对同一个或几个主帖大量回帖，导致这些用户在行为上表现出很高的协作性。

为了便于描述用户的这种协作性，可以使用“用户-主帖”网络模型来表示。该网络包含两种类型的节点：用户和主帖，这里的用户表示论坛中的一个账号，主帖表示用户为了发起新的话题而发表的帖子，也称为根帖；将用户针对主帖发表的回复帖称为回帖。图 6-13（a）是 1 个包含 6 个用户、8 个主帖的“用户-主帖”网络模型，图中圆圈表示用户，正方形表示主帖，用户和主帖之间的连边表示回复关系，例如，用户 a 和主帖 2 之间的连边表示用户 a 回复过主帖 2。

用户 a 的邻节点集合定义为与节点 a 相邻的主帖节点集合，即用户 a 回复过的主帖集合，记作 Γ_a。用户 a 和用户 b 的协作性定义为用户 a 和用户 b 的邻节点集合的相似度，相似度用 Jaccard 系数来度量，即：

$$S_{a,b}=\frac{|\Gamma_a\cap\Gamma_b|}{|\Gamma_a\cup\Gamma_b|} \tag{6-33}$$

式中，Γ_a 和 Γ_b 分别表示用户 a 和用户 b 的邻节点集合。很明显，对于任意 a 和 b，都有 $S_{a,b}=S_{b,a}$，且 $0\leqslant S_{a,b}\leqslant 1$。

（2）用户协作网络构建

论坛用户回复行为随机性大，具有很高的异质性。如果两个或多个用户表现出很高的协作性，则有理由怀疑其为网络水军账号。在“用户-主帖”网络模型的基础上，构建单日用户协作网络，对该网络的聚类特性进行分析，确定高可疑时段。

用户协作网络的构建方法如下：将用户抽象为节点，如果两个用户的协作性大于 0，即他们均回复过至少同一个主帖，则在这两个用户之间建立连边，边的权值定义为两个用户的协作性。图 6-13（b）是根据图 6-13（a）构建的用户协作网络。从图 6-13（b）可以看出，用户 a、b 和 c 之间的协作性为 1，即他们的回复对象完全相同，高度可疑。

为了更清楚地观察节点间的协作性，快速确定高可疑用户，按照边的权值对用户协作网络进行删减，仅保留其协作性大于一定阈值的边。如果仅保留图 6-13（b）中权值大于 1/3 的边，则得到图 6-13（c）。协作性高的用户表现出明显的社团特性，将此类用户看作高可疑用户。

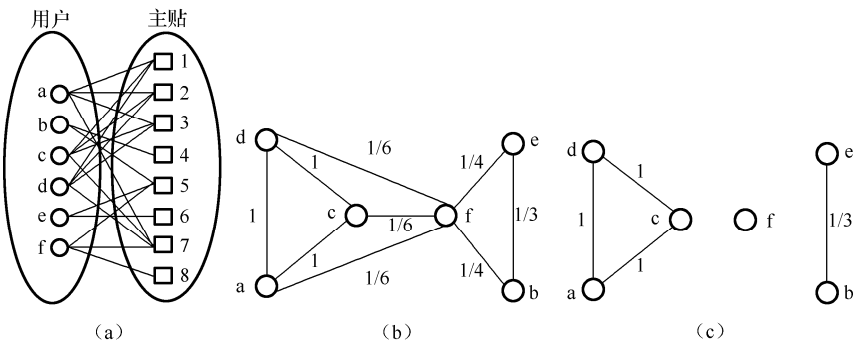


图 6-13 “用户-主帖”网络和用户协作网络模型示例

4．高可疑用户回复行为

人类打电话行为在时间上具有一定的规律性，工作时段活跃度高，休息时段活跃度低，网民回帖行为也具有类似特性。因此采用时间特征分析方法，对用户回帖行为时间特征进行分析，判定某天是否发生了网络炒作。对于确定发生了网络炒作的时段，根据网络水军相互协同这一特征推断出以“簇”形式出现的论坛用户即为网络水军账号，实施同一网络炒作的水军账号形成了水军军团，同一簇内用户共同回复的话题即为网络炒作的内容。

6.5.2 算法验证

下面通过实验数据对网络水军账号检测算法性能进行测试和验证。

1．实验数据集

实验数据来源于“新浪网－娱乐论坛－影视世界版块－影行天下子版块”2010 全年的发帖、回帖和用户信息。用 post、reply 和 user 三个数据表来存储采集到的数据，其中 post 表存储主帖信息，包括主帖 ID、发帖时间、发帖用户 ID、标题、内容；reply 表存储回帖信息，包括回帖用户 ID、回帖时间、回帖内容、对应主帖 ID；user 表存储相关用户信息，包括用户 ID、用户名、用户级别、在线时间、注册时间。

数据集共包含 4407 个主帖、80990 个回帖和 13099 个用户，其中发表过主帖的用户 1011 个，发表过回帖的用户 12929 个，2010 年全年没有发帖或回帖的用户排除在外。

2．算法有效性验证

（1）可疑时段计算

按照公式（6-26）到公式（6-32）对数据集进行统计分析，并计算三项指标的最小值、最大值及均值，计算结果如表 6-4 所示。

表 6-4　三项统计指标的计算结果

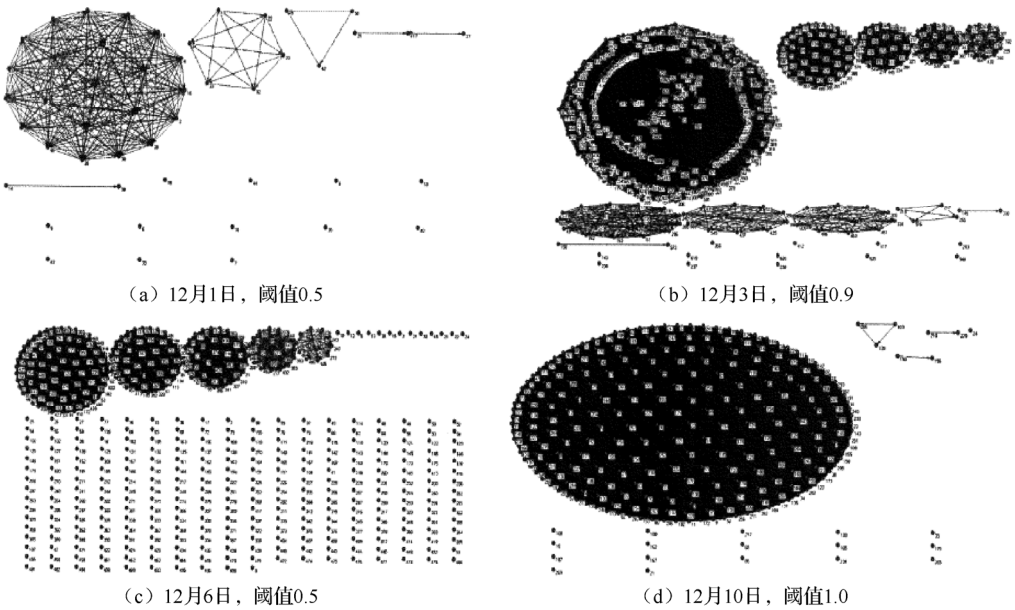
指标	统计量			
	最小值	最大值	均值	>A
RN	7	18.824	221	69
ARNU	1	29.41	2.15	103
ARNP	1	896.38	9.85	58

注：>A 表示统计指标大于其均值的天数

由表 6-4 可知，三项统计指标的异质性均非常强，大多数日子的取值都比较小。统计发现单日回帖数不小于均值的共 69 天，单日人均回帖数不小于均值的共 103 天，单日帖均回复数不小于均值的共 58 天，同时满足三个条件的共 45 天，即为可疑时段 S。

（2）高可疑时段确定

通过逐天分析可疑时段的用户回复模式，发现有 29 天的用户协作网络发生了明显聚类现象，将其确定为高可疑时段。图 6-14 显示了其中 4 天的用户协作网络，从图 6-14 可以看出，4 天用户回复行为均表现出极高的协作性。图 6-14（b）显示了 12 月 3 日仅保留权值大于 0.9 边的用户协作网络，除零星用户处于离散状态外，其他用户都聚集成为 8 个簇，同一簇内的用户协作性高达 0.9，即回复对象非常接近，高度可疑。



（a）12月1日，阈值0.5　（b）12月3日，阈值0.9

（c）12月6日，阈值0.5　（d）12月10日，阈值1.0

图 6-14　高可疑时段用户协作网络示例

（3）网络水军账号确定

为了确认高度可疑的 29 天中形成簇的用户是否为网络水军，逐天分析这些用户的回帖时间分布，统计分析结果发现，其中 7 天的用户回帖时间分布严重偏离正常用户的回帖时间分布，由此可断定这 7 天的网络论坛发生了网络炒作，它们分别是 12 月 2 日、12 月 3 日、12 月 5 日、12 月 6 日、12 月 10 日、12 月 12 日和 12 月 13 日。图 6-15 是用户回帖时间模式比较，显示了 2010 年全年及 12 月 3 日、12 月 6 日和 12 月 10 日的回帖时间在一天中的分布，其中横坐标为时间，纵坐标为该段时间的回帖数。为了便于显示，将 12 月 3 日、12 月 6 日和 12 月 10 日的统计数据分别扩大 2 倍、10 倍和 10 倍。

从图 6-15 可以看出，2010 年全年的零点回帖数较低，之后逐渐下降，并在 7 点达到谷底，这段时间正好对应人们的休息时间。之后回帖数快速上升，9 点至 23 点之间回帖数都保持在 3500 以上，其中 9 点到 18 点的回帖数略高于 18 点之后。统计结果与人们的作息规律非常吻合，也与人类打电话时间模式相一致。

观察 12 月 3 日的回帖模式，发现零点回帖数很大，之后的 5 个小时持续攀升，并在 4 点和 5 点达到最高峰；之后快速下降，9 点至 12 点回帖数均低于当天零点；13 点至 20 点，回帖数稳定在 500 左右，不到零点时的一半；之后继续下降，直到 23 点回帖量达到最低值。可以看出，12 月 3 日的用户回帖时间分布与人类作息时间完全违背。12 月 6 日的回帖时间分布与 12 月 3 日几乎相同，12 月 10 日的回帖模式与 12 月 3 日、12 月 6 日虽然不

同，但表现出异乎寻常的稳定性，也不符合人类作息规律。采用同样方法分析另外 4 天的用户回帖时间模式，发现其用户回帖时间模式也明显偏离正常用户行为特征。

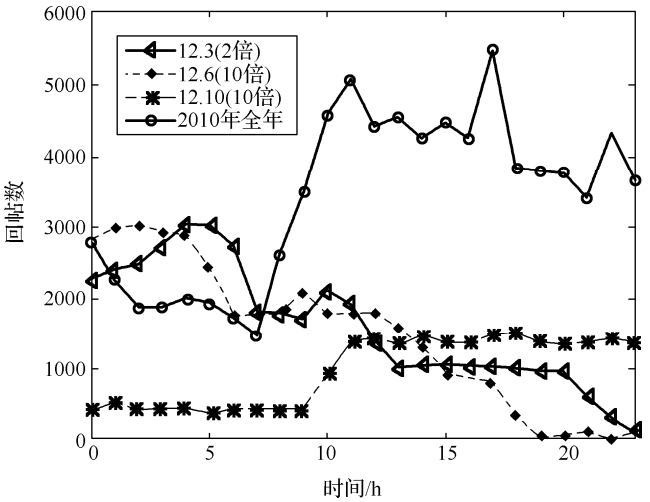


图 6-15　用户回帖时间模式比较

通过对发生网络炒作的 7 天的用户协作网络的统计分析，发现簇内共包含不同账号 556 个，构成了 1 个网络水军军团，炒作内容为当时即将上映的电影《赵氏孤儿》。

采用人工分析方式，对算法检测出的网络水军账号逐个进行分析，发现它们均为网络水军账号，算法的准确率达 100%。除算法发现的网络水军账号外，没有发现其他的可疑账号，因此该算法的漏报率为 0。

综上所述，该算法将人类行为统计分析、社会网络分析和时间特征分析等方法结合起来，逐步排除正常用户和数据，不断缩小计算范围，最终识别出网络水军账号，具有准确率高、计算量小、运算速度快等特点。

参考文献

[1] Sznajd-Weron K, Sznajd J.Opinion evolution in closed community[J].International Journal of Modern Physics C,2000,11(6):1157-1165.

[2] Pan Z. Opinions and networks: how do they effect each other[J].Computational Economics, 2012, 39(2): 157-171.

[3] Z. Pan. Trust, influence, and convergence of behavior in social networks[J]. Mathematical SocialSciences, 2010, 60(1): 69-78.

[4] Slanina F. Dynamical phase transitions in Hegselmann-Krause model of opinion dynamics and consensus[J]. The European Physical Journal B, 2011, 79(1): 99-106.

[5] Shin Y, Gupta M, Myers S. Prevalence and mitigation of forum spamming[C]//INFOCOM, 2011 Proceedings IEEE. IEEE, 2011: 2309-2317.

[6] 李雯静,许鑫, 陈正权.网络舆情指标体系设计与分析[J].情报科学, 2009, 27(7): 986-991.

[7] Chen C, Wu K, Srinivasan V, et al. Battling the internet water army: Detection of hidden paid posters[J]. arXiv preprint arXiv:1111.4297, 2011.

[8] 陈桂茸, 蔡皖东等. 一种网络论坛水军账号快速检测算法[J]. 湖南大学学报（自然科学版）, Vol.42, No.4, 2015.4.

反侵权盗版声明